View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

© 2015 David James Bloom

SENSORY DISCRIMINATION TESTING METHODOLOGY SELECTION BASED ON BEVERAGE COMPLEXITY

BY

DAVID JAMES BLOOM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Food Science and Human Nutrition
with a concentration in Food Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Shelly Schmidt, Chair
Professor Soo-Yeun Lee, Director of Research
Assistant Professor Youngsoo Lee
Associate Professor Michael J. Miller
Dr. Hwa-Young Baik, Dr Pepper Snapple Group

ABSTRACT

Sensory discrimination testing is a vital tool used by sensory professionals within the food industry. While many testing methods are available for selection, various methods have proven to differ in power. By utilizing a more powerful method, sensory professionals can limit the resources needed for testing and increase their ability to find significant differences between products. The inherent variability of food and beverage systems in addition to multidimensional changes often made during reformulation of food products can make the task of method selection challenging. The goal of this study was to determine the optimal sensory discrimination methodologies for use with multidimensional beverage systems with confusable difference comparison.

Complex beverage systems with product variations relevant to actual testing within the food industry were utilized to compare experimental results to those expected from theoretical modeling predictions. The findings from this study led to the exploration of Researcher-Designated and Panelist-Articulated specific discrimination methods to determine which procedure resulted in greater power for use in comparison of complex beverage systems. Once a more powerful specific testing procedure was established, it was then used in product comparisons with non-specific methods to determine how the degree of difference between samples influenced complex beverage systems in one-dimensional or multidimensional formulation changes. A model beverage system was created which could be altered in formulation to create multidimensional changes between samples and alter d'. The model beverage was utilized to study the impact of sample dimensionality and d' for both specified and non-specified methods. Differences in the proportion of correct response, method power, and overdispersion were observed between methods when tested with the model beverage system.

The findings from these studies emphasize the need for sensory professional to look beyond the categorization of methods by specified versus non-specified. Findings suggest methods should instead be fit to the range of d' and multidimensional makeup relevant to the samples used in testing.

ACKNOWLEDGEMENTS

Thank you to my advisor Dr. Soo-Yeun Lee. Your guidance has been invaluable not only in the completion of my studies but also in my growth professionally and personally. Thank you to my advisory committee Dr. Shelly Schmidt, Dr. Michael Miller, Dr. Youngsoo Lee, and Dr. Hwa-Young Baik. I had the great opportunity to have many of you as mentors in both undergraduate and graduate studies. You shaped the way I look at food, science, and research. I am eternally grateful for your support.

My time at the University of Illinois has been filled with a tremendous amount of love and happiness stemming from those around me. Without love from family and friends I would not be where I am today. Thank you to my mother, father, and sister for your encouragement and support. Thank you Ginn. You are a blessing in my life and words cannot encompass how much your support has meant in this process. Terri Cummings and Dawn Bohn, you both came into my life at the beginning of this adventure. You have gone above and beyond for me and I am proud to have you in my life. Thank you to the members of the Lee lab group both past and present.

There clearly is not enough room in this document to thank everyone who has provided support during my time here at the University of Illinois. You know who you are. Thank you.

Table of Contents

Chapter 1: Introduction

1.1 Research Rationale and Significance

Professionals within the food industry use discrimination testing to optimize cost, investigate customer complaints, determine shelf life, and qualify standards for use in other sensory methods. Literature has proven that the methodology used to perform sensory discrimination testing can greatly impact subject performance and test power (Byer and Abrams 1953; Ennis 1990; Ennis 1993; Bi and Ennis 1999; Angulo and others 2007; McClure and Lawless 2010). By increasing subject performance in discrimination testing, resources such as the number of subjects needed on a test, the amount of samples which need to be prepared, and the amount of time and other resources required to run the test can be greatly reduced; thus, saving time and money.

While research has been conducted in the area of methodology comparison, little research has focused on the impact of samples which differ in more than one sensory dimension on proportion of correct responses and power of each method. Additionally, the research that has been conducted within discrimination testing methodology advancement has typically been performed using samples with a large degree of difference between the two samples in comparison. When translated to the value of d', which is defined as the difference between the means of the two sample distributions represented in units of standard deviation, the large degree of difference found in the samples tested in the literature would be in the realm of 1.5. These d' values tend to be beyond those which would typically be perceived as confusable (d' ≈ 0.5 – 1.0), and are, thus, samples in which a sensory scientist within the food industry would not deem suitable for discrimination testing. Discrimination testing aims to identify if there is a

perceivable difference between samples. Samples which are not confusable will obviously be found to be different when analyzed using a discrimination test.

The current research expands upon the knowledge available in the area of discrimination testing methodologies and present findings relevant to the types of samples and number of subjects commonly used in the context of commercial food products. The research provides basis for methodology selection to sensory professionals and provides insight as to how sample complexity and degree of difference impact the expected results.

1.2 Overall Goal and Central Hypothesis

The *overall goal* of this study was to determine the optimal sensory discrimination methodologies for use with multidimensional, confusable beverage systems. Utilizing the measure of d', the perceptual distance between means of two normal distributions measured in standard deviations (O'Mahony 1992), the confusability of samples was estimated. Much of the current research available in literature has focused on sample comparisons with d' values above 1.5. In addition, the samples were either simple model solutions or samples that vary in only one dimension, which would not be typically found in commercial food products. Through the use of multidimensional samples with d' values which would lend to being confusable ($d' \leq 1$), the current knowledge of sensory discrimination methods has been expanded to more accurately reflect the true nature of sensory testing within the food industry.

Thurstonian modeling has been applied to sensory discrimination testing theory in order to describe differences observed between the results of different test methods (Frijters 1979a). Through the model's application to discrimination testing, comparison of test method power has provided guidance to sensory professionals in selecting test methods for use in the food industry. One assumption made in Thurstonian modeling is samples used in testing differ along a

unidimensional sensory attribute (Frijters 1979b). The impact of sample multidimensionality has not been fully researched in order to provide guidance on the selection of powerful test methods using complex samples.

It was hypothesized that multidimensional samples with low degree of difference between samples ($d' \leq 1$) deviate from Thurstonian modeling predictions for sensory discrimination methods.

1.3 Outline of Thesis

Chapter 2 contains a review of the current literature available on discrimination testing. Foci of the review include: an overview of commonly utilized discrimination test methods in the field of sensory science, model theories used in the identification of differences between methods, methods used to increase subject performance, samples typically utilized in research involving discrimination testing, and gaps in the current knowledge base.

Chapter 3 identifies discrimination testing methods which deviate from Thurstonian modeling predictions when multidimensional samples are utilized. It was hypothesized that specified discrimination methods would not have greater power than non-specified methods as theorized by Thurstonian modeling when multidimensional samples were utilized. To test the hypothesis, subjects were recruited to perform seven common discrimination test methods. Product categories of carbonated beverages, tea, and juice were evaluated for each test method and subject performance and test power were analyzed.

Chapter 4 compares the proportion of correct responses between Panelist-Articulated specified difference test and Researcher-Designated specified difference test when using complex samples. It was hypothesized that when samples used in discrimination testing were

multidimensional with complex formulation changes, panelists would have an increased proportion of correct responses when allowed to articulate the nature of the difference between samples than when the researcher designates the nature of the difference. Researcher-Designated methods with and without warm-up were compared to Panelist-Articulated methods using citrus-flavored carbonated beverages

Chapter 5 compares the proportion of correct responses and power of Researcher-Designated specified difference test to non-specified discrimination testing methods using multidimensional samples. It was hypothesized that differences in proportion of correct responses between Researcher-Designated specified difference test and non-specified difference tests will be decreased when tests are performed with multidimensional samples. Tea, juice, and carbonated beverages were used as samples to compare triangle, 3-AFC, tetrad, and specified tetrad methods.

Chapter 6 quantifies the level of influence sample dimensionality has on the reduction of power for specified and non-specified discrimination test methods. Two experiments were conducted to assess the influence of dimensionality. In Experiment I, it was hypothesized that multidimensionality of samples decreases proportion of correct responses on specified methods compared to non-specified methods. In Experiment II, it was hypothesized that multidimensional differences between samples created through dilution would follow basic Thurstonian modeling predictions and those created though a compensation approach would not.

Chapter 7 assesses the impact of sample d' and dimensionality on the proportion of correct responses of specified and non-specified discrimination methods in a model system. It

was hypothesized that a combination of sample dimensionality and lower d' of samples would produce results which violate Thurstonian model predictions in a model beverage.

Chapter 8 concludes the dissertation by exploring future directions for the research. Additional consumer testing research focusing on environmental and sample contextual impacts on sensory consumer testing is discussed. An additional proposal to expand the research to investigate the effects of the degree of difference between samples to sensory consumer testing is included.

1.4 References

Angulo O, Lee H, O'Mahony M. 2007. Sensory difference tests: Overdispersion and warm-up. Food Quality and Preference 18(2):190-5.

Bi J, Ennis DM. 1999. The Power Of Sensory Discrimination Methods Used In Replicated Difference And Preference Tests. J.Sens.Stud. 14(3):289-302.

Byer AJ, Abrams D. 1953. A comparison of the triangular and two-sample taste-test methods. Wallerstein Lab Commun 16((54)):253-60.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Frijters JER. 1979a. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER. 1979b. Variations of the triangular method and the relationship of its unidimensional probabilistic models to three-alternative forced-choice signal detection theory models. Br.J.Math.Stat.Psychol. 32(2):229-41.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

O'Mahony M. 1992. Understanding Discrimination Tests: A User-friendly Treatment Of Response Bias, Rating And Ranking R-index Tests And Their Relationship To Signal Detection. J.Sens.Stud. 7(1):1-47.

Chapter 2: Literature Review

2.1 Introduction

Sensory discrimination testing at its most basic level is performed to identify if a difference can be perceived between two samples (Stone and others 2012). Using the results of discrimination testing, sensory professionals can guide product developers in the creation of samples which may differ in formulation but do not significantly differ in sensory perception. Discrimination testing can be used for many reasons including formulation changes, processing alterations, or even to support claims where difference is desired such as "new and improved" (Lawless and Heymann 2010). Various methods are available for sensory professionals to select; some with a history of use dating back decades (Brandt and Arnold 1977).

While there is a history of utilizing discrimination testing in the food industry, there is ongoing debate as to which method should be employed in testing procedures (Ennis 2012). Conducting data analysis beyond simple binomial statistics has led to the comparison of testing procedures for suitability and effectiveness within the food industry. Comparison of discrimination testing methods has largely been completed using simple test samples which often have a large degree of sensory difference between them (Ishii and others 2014a). While models exist to address the impacts of sample dimensionality on testing results (Ennis and Mullen 1986a), there is little published experimental data to compare with multidimensional models to determine model validity. The current review will explore literature available on comparison of methods and address areas where additional research is needed.

2.2 Discrimination Testing Methods

2.2.1 Non-Specified Test Methods

There are several difference tests available that do not require specification of an attribute with which subjects will differentiate samples. These methods include the same-different method (Peryam and others 1954), duo-trio (Dawson and others 1951), triangle (Helm and Trolle 1946), and tetrad (Lockhart 1951). Details about each method can be found in Table 2.1. Non-specified test methods have broad use within the food industry and are useful when changes made between samples do not lead to a clear attribute by which samples differ (Bi 2008).

Table 2.1 Description of commonly used non-specified discrimination testing methods.

Test Method	Task	Sample Presentation Orders	Chance Probability
Same-Different	Identify if the samples are the same or different	AA, BB, AB, BA	1/2
Duo-trio	Identify the sample that is the same as the reference	RA:AB, RA:BA, RB:BA, RB:AB	1/2
Triangle	Identify the odd sample	AAB, ABA, BAA, BBA, BAB, ABB	1/3
Tetrad	Group the samples into two groups of two based on similarity	AABB, ABAB, BBAA, BABA	1/3

Historically, the triangle test method has been one of the most popular sensory methods utilized within the food industry. In fact, 25 years after its introduction into the food industry, the triangle test was the most commonly used sensory method for major food companies in the United States (Brandt and Arnold 1977). Ease of use and early adoption may be a reason why the triangle test method is still commonly used today (Plotto and others 2010; Jung and others 2010; Garcia and others 2012; Marconi and others 2014; Pellegrino and others 2015).

Recently, the triangle test has begun to lose favor within the food industry and the tetrad method has become a more broadly utilized testing method for non-specified discrimination testing (Masuoka and others 1995; Delwiche and O'Mahony 1996; Ennis and Jesionka 2011; Ennis 2012; Rousseau and Ennis 2013; O'Mahony 2013; Garcia and others 2013; Bi and O'Mahony 2013; Ennis and Christensen 2014; Ishii and others 2014a; Ishii and others 2014b; Ennis and Christensen 2015; Xia and others 2015). The increased attention given to the tetrad method is the result of the method being found to have greater power than the triangle method (Ennis 2012). Advances in how test methods are compared, discussed in a later section, have led to increased discussion and scrutiny of discrimination methods. The goal of the attention paid to the tetrad method in recent years is the possibility of identifying a new test method capable of broad use with minimal resource input.

2.2.2 Specified Test Methods

Specified discrimination testing methods utilize a term given to subjects during testing procedures which indicates how samples differ. By providing an attribute to subjects, performance on specified test methods increases compared to similar non-specified test methods (Byer and Abrams 1953; Gridgeman 1970; Frijters and others 1980; Ennis 1990; Ennis 1993; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005).

Specified discrimination testing methods typically differ based on the number of samples presented during testing. Common specified discrimination testing procedures include the 2-Alternative Forced Choice (2-AFC) method (Green and Swets 1966), 3-AFC method (Green and

Swets 1966) and the specified tetrad (Ennis and Jesionka 2011). More details about each method can be found in table 2.2.

Table 2.2 Description of commonly used specified discrimination testing methods. Sweeter is listed as an example of possible terms used in specified methods.

Test Method	Task	Sample Presentation Orders	Chance Probability
2-AFC	Identify the "sweeter" sample	AB, BA	1/2
3-AFC	Identify the "sweeter" sample	AAB, ABA, BAA	1/3
Specified Tetrad	Identify the two samples that are "sweeter"	AABB, ABAB, BBAA, BABA	1/6

2.2.3 Panelist Articulated Specified Methods

Specified discrimination testing methods require researchers to designate to subjects how samples differ. As reformulation of a product often involves changing more than one ingredient, it may be difficult or impossible to provide subjects with an attribute with which to differentiate samples. Thieme and O'Mahony (1990) have suggested the use of subjects to articulate differences between samples as a way of overcoming this limitation of specified discrimination tests. By using subjects to articulate a difference between samples, subjects create an attribute specific to their sensory perception and transfer this attribute into the testing procedures.

Following these suggestions, several studies have been conducted to compare methods using panelist articulated differences to more traditional methods. McClure and Lawless (2010) utilized panelists to articulate differences between samples prior to completing 2-AFC testing. Although specified procedures were observed to have higher proportion of correct responses compared to triangle procedures, they had significantly lower d' values (discussed in the next section) than triangle procedures. Another issue discovered in the study was the attenuation of

subjects to irrelevant attributes which lowered performance for the test methods where subjects articulated differences between samples. Xia and others (2015) conducted testing using panelists to articulate differences between samples for 2-AFC and 3-AFC procedures. In addition to describing the difference between samples, subjects were also asked to designate a preference between samples which the authors believed to be more stable than difference alone. These panelist articulated procedures were found to have higher proportions of correct responses than triangle and tetrad methods.

2.3 Model Theories

Discrimination testing methods can differ based on the number of samples used in testing, the chance probability of selecting a correct response, and what question subjects are being asked. These difference in methods can influence test performance and lead to conflicting results based on which method is selected for testing. Most notably within sensory testing is the comparison of the 3-AFC and triangle test.

First discovered by Byer and Abrams (1953), subjects demonstrated greater performance on 3-AFC procedures than when completing the triangle test, albeit the only differences between the procedures was in the instructions given to subjects. This discrepancy between methods went on to be termed, the "paradox of discriminatory non-discriminators" (Gridgeman 1970). The 3-AFC method has since been shown to result in a larger proportion of correct responses than the triangle test method in several studies (Frijters and others 1980; Ennis 1990; Ennis 1993; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005). Resolving the "paradox of discriminatory non-discriminators" was achieved by exploring the decision strategies induced

through differences in instructions for the 3-AFC and triangle test (Frijters 1979a). The models created to compare discrimination testing methods are expanded upon below.

2.3.1 Guessing Model

Data collected from sensory discrimination testing falls under the category of binomial data (O'Mahony 1986). Results fall into two categories, correct responses and incorrect responses. While the results of discrimination testing may be simple to categorize, these categories may not fully reflect the true population of subjects who are able to discriminate between samples. Forced choice discrimination testing methods have a probability of subjects obtaining a correct response purely by chance. How guessing impacts performance on discrimination tests is addressed using the guessing model (Lawless and Heymann 1998).

When viewing the results from a discrimination test, the group of subjects who obtain a correct response include discriminators who perceive differences between samples and non-discriminators who guess correctly (Lawless and Heymann 1998). The proportion of non-discriminators who guess correctly can be subtracted from the proportion of correct responses to provide a greater estimate of the true proportion of discriminators using Abbot's formula (Morrison 1978). Determining the proportion of probable discriminators allows researchers to determine significance for testing, which is not reliant on the number of subjects used to conduct the test (Lawless and Heymann 1998).

While determining a proportion of probable discriminators is an added benefit of the guessing model compared to simple binomial analysis, it does not resolve the issue surrounding the "paradox of discriminatory non-discriminators." The proportion of correct responses is method dependent (Bi 2008). In order to compare different discrimination testing methods to one another and resolve the "paradox of discriminatory non-discriminators," a different approach was

needed. As a result Thurstonian modeling is often used in the discussion of discrimination testing method comparison.

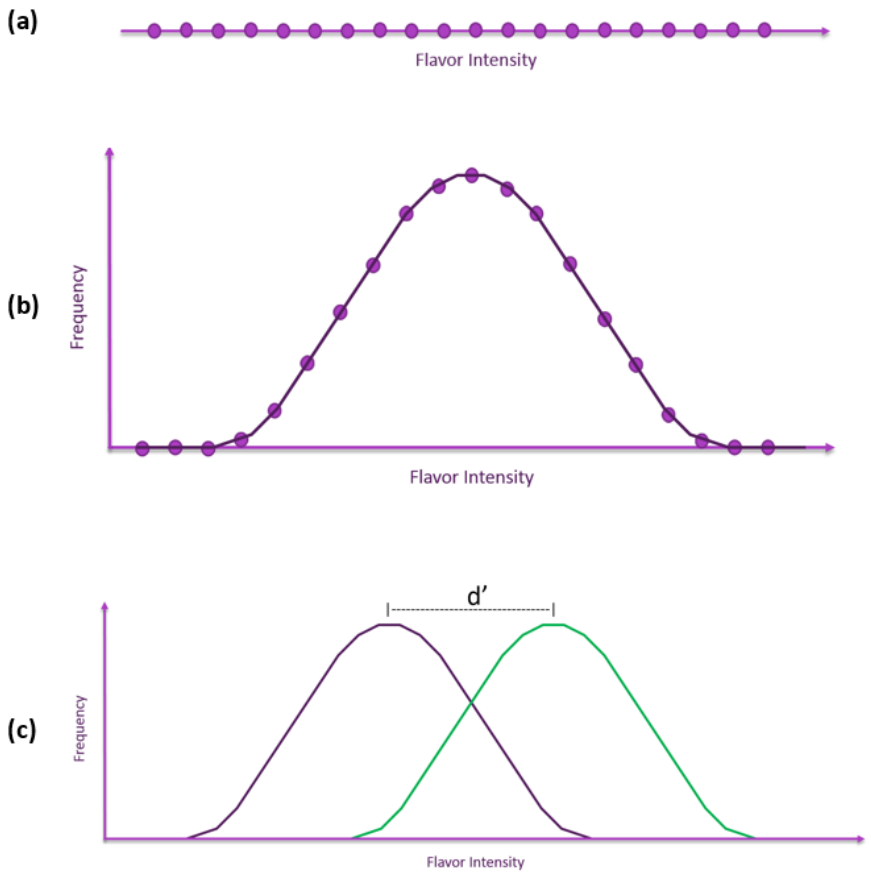
2.3.2 Thurstonian Model

The Law of Comparative Judgement was first described by L.L Thurstone in 1927 (Thurstone 1994). In this law, Thurstone described how judgements made about a stimulus were not constant over time. These judgements are believed to fall along a psychological continuum that deviate in a way which are thought to be Gaussian in distribution. Thurstone's Law of Comparative Judgement was utilized by Frijters (1979a) in application to sensory discrimination testing. The work of Thurstone and Frijters has led to an increased understanding of how subjects discriminate food and beverage samples in sensory testing and have aided in the resolution of the "paradox of discriminatory non-discriminators" (Byer and Abrams 1953; Gridgeman 1970).

A more detailed explanation of how Thurstonian modeling has been applied to sensory discrimination testing can be performed with the help of Figure 2.1 derived from O'Mahony and others (1994). The axis used in this example could be the intensity of perception of any attribute in question, but the example used here is flavor. Upon repeated tastings of a stimulus, flavor intensity will vary over time for an individual subject along a continuum (Figure 2.1a). As seen in Figure 2.1b, when a frequency distribution is created for the intensity ratings of the sample, the resulting distribution is assumed to be normal. When two stimuli are perceived, two distributions are created. The assumptions made by the models describing these distributions are that the two distributions are normal and have equal variance and the samples from which the distributions are created vary along a unidimensional sensory attribute (Frijters 1979b). If the

distributions overlap as in Figure 2.1c, the stimuli are considered to be confusable (O'Mahony and others 1994).

Figure 2.1. Perception of sensory stimuli as explained using Thurstonian modeling. Sensory perception of a single stimuli is not constant from trial to trial and the frequency of perceptions approach a normal distribution. Derived from (O'Mahony and others 1994).

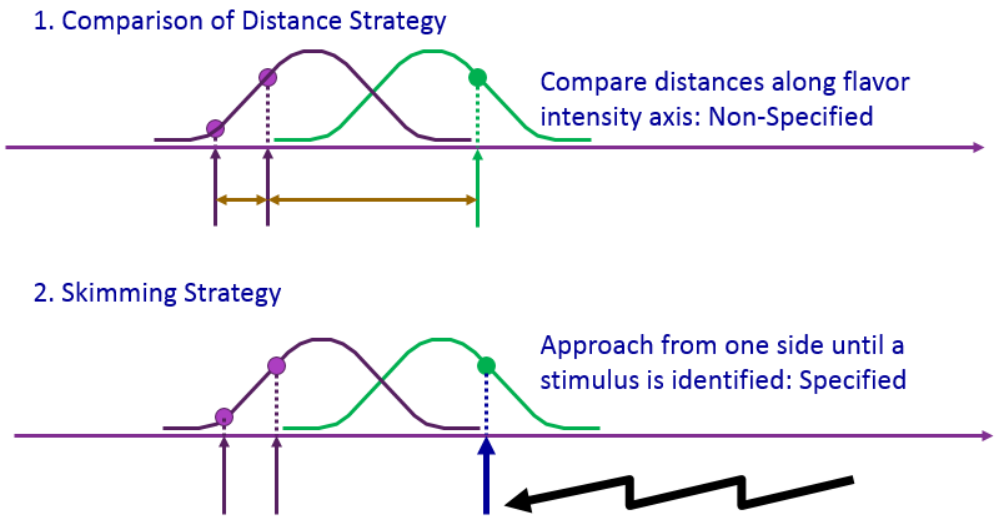


Borrowed from signal detection theory, d' (Figure 2.1c) is an estimate of judges' sensitivity. The d' of a sample set is utilized to compare results from different methods (O'Mahony 1992). While the proportion of correct responses may change from one method to another, it is believe that the sensitivity of the judges measured by d' remain s constant. We utilize this measure of sensitivity, d', as a way of determining how far away two samples may be in terms of perceptual distance. Samples with a higher d' have a greater distance between

perceptual means than samples with low d' (O'Mahony and others 1994), and would thus be more distinguishable from one another.

To resolve the paradox of discriminatory non-discriminators, Frijters utilized the theories set in place by Thursone and expanded upon them to explain how some methods resulted in a larger proportion of correct responses than other methods. Although the setup of the test is nearly identical in both 3-AFC and triangle procedures, the instructions given to subject are different. The differences in instructions between triangle and 3-AFC methods are believed to change the decision strategy used in completing the different methods (Frijters 1979a).

Figure 2.2. Depiction of cognitive strategies believe to be utilized in specific and non-specific discrimination testing. Derived from (O'Mahony and others 1994).



When subjects approach a triangle test they are thought to take on a "comparison of distances" strategy to complete the test (Figure 2.2). Within the comparison of distances strategy subjects will analyze each sample. Two samples will be drawn from the first distribution and one sample from the second distribution. The distances along the sensory continuum between

samples will be compared. The samples with the smallest distance between one another will be deemed similar and the sample with the largest distance away from the other samples will be deemed different (O'Mahony and others 1994).

When subjects approach a 3-AFC test they are instructed to focus on either the strongest or weakest sample. These instructions are believed to illicit a "skimming" strategy (Figure 2.2). As in the triangle test, two samples will be drawn from the first distribution and one from the second distribution. Instead of comparing the distances between samples, subjects simply approach from either extreme of the sample set and choose either the strongest perception or the weakest perception depending on instructions.

Frijters borrowed models created to represent both the comparison of distances strategy (Ura 1960) and skimming strategy (Green and Swets 1966) to resolve the "paradox of discriminatory non-discriminators." Since d' remains constant between methods, it can be used in the comparison of methods that may apply difference decision strategies. When comparing d' values from the work of Byer and Abrams (1953), Frijters found that while the proportion of correct responses for the two methods differed, d' remained constant thus resolving the "paradox of discriminatory non-discriminators."

Since resolving the paradox, Thurstonian modeling has been widely used in the analysis of sensory samples. Studies utilizing the models discussed will be explored in further detail in a future section.

2.3.3 Multidimensional Thurstonian Model

One limitation of traditional Thurstonian modeling is the assumption that samples differ along one dimension. Samples used by the food and beverage industries in discrimination testing

do not always vary along just one sensory dimension. The use of multidimensional models in sensory testing, derived by Ennis and Mullen (1985), have been offered as a way of removing the assumption of unidimensionality.

Ennis and Mullen focused on the impact of increasing dimensionality on the resulting proportion of correct responses using model simulations for the triangle test (Ennis and Mullen 1985; Ennis and Mullen 1986a) and later the duo-trio test method (Ennis and Mullen 1986b; Mullen and Ennis 1987). Using the models, Ennis and Mullen demonstrated that by increasing dimensionality of samples, a reduction in the proportion of correct responses may occur. Additionally, the increase of dimensionality will decrease the power of a test method (Ennis and Mullen 1986a).

The decrease in test power with increasing dimensionality was observed for samples, in which samples differed by only one dimension and assuming that variables were not correlated. When multiple dimensions change between samples or when correlation of variables exist, proportion of correct responses for a test method will not only depend on the number of dimensions present, but also on the degree of correlation of dimensions and the direction of difference between samples (Ennis and Mullen 1986a). The fact that many sample factors may impact the results of discrimination testing requires extensive knowledge of samples used in testing in order to properly interpret results.

Multidimensional modeling of discrimination testing has shown that sample formulation and sensory perception increases the need for understanding samples beyond d'. Instead of focusing on the method that results in a larger proportion of correct responses at a certain d', it may be best to compare performance of methods with samples fit to a company's or research facility's product focus. The need for comparing methods beyond d' becomes all the more

important when it is understood that one discrimination testing method may be more sensitive than another to changes made in a multidimensional product (Ennis and Mullen 1986b).

2.4 Factors Influencing Method Power

2.4.1 Test Method

Ennis and Jesionka (2011) defined discrimination test power as "the capacity of that test to reliably detect differences between products". When conducting a statistical test, power is determined to be 1-β (Schlich 1993). As β is the probability of missing a true difference, test methods with low power have a higher risk of missing true differences between products than test methods with higher power. Selecting a test method with high power is advantageous as it would require a smaller sample size than lower power methods for equivalent risk levels (Rousseau 2003). Powerful test methods can reduce resources needed to complete testing both in terms of time and physical resources such as samples and compensation.

As discussed previously, the decision strategy utilized by different test methods impact performance and ultimately power of test methods. McClure and Lawless (2010), using tables published by Ennis (1993), demonstrated the impact of test power on the number of subjects needed to complete a triangle test with the same power as the 2-AFC method. At an alpha risk of 5% and a beta risk of 20%, in order to identify a difference between samples with a d' of 1.0, only 20 subjects would be required for 2-AFC, while 197 subjects are needed for the triangle test. At lower levels of d', the subjects needed for the triangle test increases to nearly 35 times more compared to the 2-AFC. As the models predicting these figures are based on Thurstonian estimations, experimental data using complex samples is needed to determine if differences in the number of subjects needed for high power hold true in actual product testing.

In addition to the number of subjects utilized in testing, the number of samples presented to subjects during testing can also impact power. It is believed more samples used in testing increases the memory load needed in completing the test. Increasing the memory load has an effect of decreasing d' and power for a test method (O'Mahony 2013). A comparison between the 2-AFC and 3-AFC methods is a prime example. While theory predictions indicate more power for the 3-AFC method, the 2-AFC has been shown to be more powerful due to lower memory load (Rousseau and O'Mahony 1997; Dessirier and O'Mahony 1998).

Memory load may again become a concern when methods such as the tetrad, which utilizes four test samples, is employed. If the d' for an unspecified tetrad test drops by one-third, it loses its theoretical power advantage over the triangle test method (Ennis 2012). Several studies have identified a reduction in d' compared to the triangle test. While remaining more powerful, the tetrad method was observed to have lower d' than the triangle test method when children completed testing with apple juice samples (Garcia and others 2012). Carlisle (2014) found the power of the tetrad method to be lower or equivalent to the triangle for several baked bean products, apple sauce, tomato sauce, and oat cereal. It is clear that the sample complexity has varying influence on the power of discrimination methods.

2.4.2 Warm-up

While the number of samples presented to subjects can impact performance, providing samples prior to testing has actually shown to improve performance during data collection. Presenting samples prior to testing to induce an increase in performance by subjects has been termed a warm-up effect. Heron (1928) demonstrated an increase in performance for the learning

of nonsense syllables with the addition of a warm-up session prior to testing. This was one of the first published discussion of a warm-up effect in literature.

Applying the findings of Heron to discrimination testing, O'Mahony and others (1988) compared triangle procedures with and without warm-up samples. To do so, in warm-up procedures, subjects were provided two different test samples and were asked to taste them back and forth until they felt they could distinguish between samples, 3-10 pairs were used. The study demonstrated an increase in performance for triangle tests with warm-up using salt water solutions and orange juices with added citric acid. Since the work of O'Mahony and others (1988), a warm-up effect has been found to increase performance in several other published studies (O'Mahony and others 1988; Thieme and O'Mahony 1990; Dacremont and others 2000; Mata-Garcia and others 2007; Angulo and others 2007).

As discussed previously, one proposed manner of utilizing a specified difference test when an attribute with which to differentiate samples is not easily identified is by way of subjects articulating the difference between the samples. As part of this procedure, subjects may be asked to taste test samples back and forth until they are able to specify a difference between the samples. These procedures mimic those used in many warm-up studies including that of O'Mahony and others (1988). It remains unclear if the samples utilized in the articulation of the difference between the two products also induce a warm-up effect. McClure and Lawless (2010) observed several subjects attenuating to irrelevant sample differences during the articulation process, which impacted performance on testing. Separating the effects of warm-up and panelist articulation process has yet to be fully explored but may serve to further understand the impact of both warm-up and panelist articulation on the power of specified methods.

2.5 Sample Dimensionality

Basic Thursonian modeling relies on the assumption that samples differ along one sensory dimension (Frijters 1979b). Research comparing sensory discrimination test methods has typically utilized samples which also differ along one sensory dimension with a few exceptions. Generally, samples which have been used in comparison studies have fallen under three main categories: 1) one-dimensional model solutions, 2) products which differ by one formulation change, or 3) multidimensional samples changes. Research within each of these categories will be discussed below.

2.5.1 Model Solution Research

Foundational research in the area of sensory discrimination testing methods was performed using basic taste solutions. Byer and Abrams (1953) work, which led to the focus of the impact sensory methodology had on subject performance, was conducted using quinine sulfate solutions and dextrose solutions. The discrepancy found in subject performance between specified and non-specified methods led to the incorporation of Thurstonian modeling into discrimination testing (Frijters 1979a). Modeling comparisons were easily fit due to the fact that samples were unidimensional in formulation.

In addition to the foundational work of method comparison, model solutions have been commonly utilized in the exploration of sample effects during testing. For example, the study of sample sequence effects on the outcomes of discrimination testing is commonly explored using sodium chloride solutions (Tedja and others 1994; Dessirier and O'Mahony 1998). The use of model solutions in conducting foundational research has led to important discoveries in the field of sensory science.

What is concerning about reliance on research using model solutions is in the application of findings to broader food industry contexts. Sample variance can greatly increase as product complexity increases. Take an example of chicken noodle soup; the presence of noodles, protein, and vegetables with the product inherently create issues when conducting method comparisons. How can one be sure that the makeup of the sample is consistent between and within testing sessions? The variability created by the sample complexity may produce unknown differences between samples used in testing which could alter subject perception and lead to unintended differences between samples. Additional studies which utilize basic model solutions in the comparison of discrimination testing procedures can be found in Table 2.3.

Table 2.3 Examples of studies focusing on sensory discrimination testing methodologies which utilize model solutions in the comparison of test methods. Citation, makeup of samples utilized in testing, and methods compared are provided.

Citation	Sample Makeup	Methods
(Byer and Abrams 1953)	Quinine sulfate solutions, dextrose solutions	Triangle, 2-AFC
(Frijters and others 1982)	NaCl solutions	Triangle
(O'Mahony and Odbert 1985)	NaCl solutions	3-AFC, Triangle, Duo-Trio
(Thieme and O'Mahony 1990)	NaCl solutions	Duo-Trio, Paired Comparison, A-Not A
(Tedja and others 1994)	NaCl solutions	Triangle, 3-AFC
(Dessirier and others 1999)	NaCl solutions	3-AFC
(Braun and others 2004)	NaCl solutions	2-AFC, 2-AC
(Lau and others 2004)	NaCl solutions	Same-Different, Triangle
(Lee and O'Mahony 2007)	NaCl solutions	3-AFC
(Angulo and others 2007)	NaCl solutions	2-AFC, 3-AFC, Duo-Trio, Triangle

2.5.2 One Dimensional Sample Changes

Studies within literature commonly utilize samples which differ in one ingredient between control and variant samples when comparing sensory discrimination test methods. As observed in Table 2.4, samples are typically fluid beverages as homogeneity of samples is easily achieved. Rousseau and O'Mahony (1997) utilized commercially available yogurt to which they added sucrose to increase sweetness between samples. Using the yogurt samples, Thurstonian modeling predictions were confirmed as the 3-AFC method resulted in a larger proportion of correct responses than the triangle test method while maintaining a similar level of d' 1.7 and 1.9.

Table 2.4 Examples of studies focusing on sensory discrimination testing methodologies which utilize multidimensional samples with one ingredient change in the comparison of test methods. Citation, makeup of samples utilized in testing, and methods compared are provided.

Citation	Samples	Methods
(O'Mahony and Goldstein 1986)	Sparkling flavored water (sodium saccharin)	Triangle
(Cubero and others 1995)	Citrus flavored beverage (sucrose)	Same-different
(Masuoka and others 1995)	Beer (isohumulone)	Triangle, 3-AFC
(Huang and Lawless 1998)	Flavored beverage (sucrose)	Paired Comparison, Triangle, ABX, 3-AFC, Dual Standard, Duo-trio
(Rousseau and O'Mahony 1997)	Yogurt (sucrose)	Triangle, Duo-Trio, Same-Different
(Rousseau and others 2002)	Orange flavored beverage (sucrose)	Duo-Trio, DTM, 2-D-AFC
(Braun and others 2004)	Sparkling mineral water (carbonation)	2-AFC, 2-AC
(Liggett and Delwiche 2005)	Flavored beverage (sucrose)	2-AFC, 3-AFC, Triangle, Duo-Trio
(McClure and Lawless 2010)	Flavored beverages (sucrose or citric acid)	Triangle, 2-AFC

The d' of samples utilized in the Rousseau and O'Mahony study is not uncommon to those found in literature despite the fact that a d' of 2.0 is a relatively high degree of difference between samples (Ishii and others 2014a). Samples with a d' of 2.0 would result in 92% of correct responses for the 2-AFC method. Differences this high may not be relevant to samples commonly tested in a discrimination test setting in the food industry as the samples are no longer confusable. Samples utilized in a study conducted by Huang and Lawless (1998) resulted in perfect discrimination for two of the methods used in testing. At this level of discrimination d' is considered infinite as distributions do not overlap. The tendency to utilize samples at low levels of confusability in addition to samples which only differ in one dimension is not typical to what may be encountered within industrial testing situations.

2.5.3 Multidimensional Sample Changes

More common to the type of samples which are believed to be used on in an industrial setting are samples with multidimensional formulation changes. To address these factors, previously conducted studies often create multidimensional changes using dilution of liquid samples. Examples of several studies using dilution can be found in Table 2.5. While dilution of samples will change multiple attributes of a product, all attributes are changing in the same direction and by the same dilution factor. It is difficult to imagine a situation in which dilution would be commonly utilized in the food industry for product development or maintenance purposes.

Table 2.5 Examples of studies focusing on sensory discrimination testing methodologies which utilize multidimensional samples complex ingredient changes in the comparison of test methods. Citation, makeup of samples utilized in testing, and methods compared are provided. Samples listing diluted used sample dilution between control and variant samples. Samples listing type differed based on the type of commercially sourced products used between sample pairs.

Citation	Samples	Methods
(O'Mahony and Goldstein 1986)	Wine (diluted)	Triangle
(Huang and Lawless 1998)	Tea (type)	Paired Comparison, Triangle, ABX, 3-AFC, Dual Standard, Duo-trio
(Rousseau and others 1999)	Mustard (type)	Same-different, Triangle
(McClure and Lawless 2010)	Broth and Tea (dilution)	Triangle, 2-AFC
(Van Hout and others 2011)	Margarine (type)	2-AFCR, A-Not A, 2-AFC
(Garcia and others 2013)	Apple juice (diluted)	Specified Tetrad, 2-AFC
(Ishii and others 2014a)	Juice (diluted)	Triangle, Tetrad

Another less commonly utilized technique used to create multidimensional changes between test samples is by using commercial products. This may be the closest to industrial testing situation of the samples found in literature. One potential issue with utilizing commercially available samples is the potential for large differences between samples. Huang and Lawless (1998) utilized commercial tea samples and observed d' values as large as 2.93 during testing. Rousseau and others (1999) utilized commercially available mustards and observed d' values as large as 2.31 during testing comparing Triangle and 2-AFC methods. Van Hout and others (2011) conducted test method comparisons using more confusable margarine samples with a d' or approximately 1.5. Although the naming of methods in the study resemble specified test, the study compared three non-specified test methods. It is unclear from the findings available in literature how multidimensionality may influence performance on specified test methods compared to non-specified methods for confusable samples.

2.6 Conclusions

The application of Thurstonian models to sensory discrimination testing has led to enhanced understanding of differences between method results. An inherent limitation to basic Thurstonian modeling is the assumption that samples differ along one sensory dimension. This limitation has led to much of the current research comparing methods to be conducted using model solutions or samples with changes in one ingredient. Differences in the proportion of correct responses using model solutions should be seen as an idealized testing situation and may not imply differences in samples with more complex changes.

Multidimensional Thurstonian modeling suggests selecting methods based on sample variation as some methods may be more responsive than others to types of sample dimensionality. There is little data available within literature using complex, confusable samples with multidimensional formulation changes. Expanding research in this area will allow for more application of Thurstonian models to products relevant to the food industry.

2.7 References

Angulo O, Lee H, O'Mahony M. 2007. Sensory difference tests: Overdispersion and warm-up. Food Quality and Preference 18(2):190-5.

Bi J. 2008. Sensory discrimination tests and measurements: Statistical principles, procedures and tables. John Wiley & Sons.

Bi J, O'Mahony M. 2013. Variance of d? for the Tetrad Test and Comparisons with Other Forced-Choice Methods. J.Sens.Stud. 28(2):91-101.

Brandt FI, Arnold RG. 1977. Sensory tests used in food product development. Food Product Development 1156.

Braun V, Rogeaux M, Schneid N, O'Mahony M, Rousseau B. 2004. Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. Food Quality and Preference 15(6):501-7.

Byer AJ, Abrams D. 1953. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Carlisle SL. 2014. Comparison of Triangle and Tetrad Discrimination Methodology in Applied, Industrial Manner. [dissertation]. Knoxville: University of Tennessee.

Cubero E, Avancini de A., Tereza C., O'Mahony M. 1995. Cognitive aspects of difference testing: Memory and interstimulus delay. J.Sens.Stud. 10(3):307-24.

Dacremont C, Sauvageot F, Duyen TH. 2000. Effect Of Assessors Expertise Level On Efficiency Of Warm-up For Triangle Tests. J.Sens.Stud. 15(2):151-62.

Dawson EH, a Redstrom R, Harris BL. 1951. Sensory methods for measuring differences in food quality. US Government Printing Office.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Dessirier J, O'Mahony M. 1998. Comparison of d′ values for the 2-AFC (paired comparison) and 3-AFC discrimination methods: Thurstonian models, sequential sensitivity analysis and power. Food Quality and Preference 10(1):51-8.

Dessirier J, Sieffermann J, O'Mahony M. 1999. Taste Discrimination By The 3-Afc Method: Testing Sensitivity Predictions Regarding Particular Tasting Sequences Based On The Sequential Sensitivity Analysis Model. J.Sens.Stud. 14(3):271-87.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM, Mullen K. 1986a. A multivariate model for discrimination methods. J.Math.Psychol. 30(2):206-19.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Ennis DM, Mullen K. 1986b. Theoretical aspects of sensory discrimination. Chemical Senses 11(4):513-22.

Ennis DM, Mullen K. 1985. The effect of dimensionality on results from the triangular method. Chemical Senses 10(4):605-8.

Ennis JM. 2012. Guiding The Switch From Triangle Testing To Tetrad Testing. J.Sens.Stud. 27(4):223-31.

Ennis JM, Christensen R. 2015. A Thurstonian comparison of the Tetrad and Degree of Difference tests. Food Quality and Preference 40, Part B263-9.

Ennis JM, Christensen RHB. 2014. Precision of measurement in Tetrad testing. Food Quality and Preference 32, Part A(0):98-106.

Ennis JM, Jesionka V. 2011. The Power Of Sensory Discrimination Methods Revisited. J.Sens.Stud. 26(5):371-82.

Frijters JER. 1979a. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER. 1979b. Variations of the triangular method and the relationship of its unidimensional probabilistic models to three-alternative forced-choice signal detection theory models. Br.J.Math.Stat.Psychol. 32(2):229-41.

Frijters JER, Blauw YH, Vermaat SH. 1982. Incidental training in the Triangular Method. Chemical Senses 7(1):63-9.

Frijters JER, Kooistra A, Vereijken PFG. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Percept.Psychophys. 27(2):176-8.

Garcia K, Ennis JM, Prinyawiwatkul W. 2013. Reconsidering the Specified Tetrad Test. J.Sens.Stud. 28(6):445-9.

Garcia K, Ennis JM, Prinyawiwatkul W. 2012. A large-scale experimental comparison of the tetrad and triangle tests in children. J.Sens.Stud. 27(4):217-22.

Green DM, Swets JA. 1966. Signal detection theory and psychophysics.

Gridgeman NT. 1970. A Reexamination of the Two-Stage Triangle Test for the Perception of Sensory Differencesa. J.Food Sci. 35(1):87-91.

Helm E, Trolle B. 1946. Selection of a taste panel. Wallerstein Lab Commun 9(28):181.

Heron WT. 1928. The Warming-Up Effect in Learning Nonsense Syllables. The Pedagogical Seminary and Journal of Genetic Psychology 35(2):219-28.

Huang Y-, Lawless HT. 1998. Sensitivity of the ABX discrimination test. J.Sens.Stud. 13(2):229-39.

Ishii R, O'Mahony M, Rousseau B. 2014a. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Qual.Preference 31(1):49-55.

Ishii R, O'Mahony M, Rousseau B. 2014b. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Quality and Preference 31(0):49-55.

Jung D, Hong J, Kim K. 2010. Sensory characteristics and consumer acceptability of beef soup with added glutathione and/or MSG. J.Food Sci. 75(1):S36-42.

Lau S, O'mahony M, Rousseau B. 2004. Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. Percept.Psychophys. 66(3):464-74.

Lawless HT, Heymann H. 2010. Sensory evaluation of food: principles and practices. Springer Science & Business Media.

Lawless HT, Heymann H. 1998. Discrimination Theories and Advanced Topics. In: R. Bloom, editor. Sensory Evaluation of Food Principles and Practices. 1st ed. Aspen Publishers, Inc. p 140.

Lee H, O'Mahony M. 2007. Difference test sensitivity: Cognitive contrast effects. J.Sens.Stud. 22(1):17-33.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Lockhart E. 1951. Binomial systems and organoleptic analysis. Food Technology 5428.

Marconi O, Martini R, Mangione A, Falconi C, Pepe C, Perretti G. 2014. Palatability and Stability of Shortbread Made with Low Saturated Fat Content. J.Food Sci. 79(4):C469-75.

Masuoka S, Hatjopoulos D, O'Mahony M. 1995. Beer Bitterness Detection: Testing Thurstonian And Sequential Sensitivity Analysis Models For Triad And Tetrad Methods. J.Sens.Stud. 10(3):295-306.

Mata-Garcia M, Angulo O, O'Mahony M. 2007. On Warm-up. J.Sens.Stud. 22(2):187-93.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

Morrison DG. 1978. A Probability Model for Forced Binary Choices. The American Statistician 32(1):23-5.

Mullen K, Ennis DM. 1987. Mathematical formulation of multivariate euclidean models for discrimination methods. Psychometrika 52(2):235-49.

O'Mahony M, Goldstein L. 1986. Effectiveness of sensory difference tests: Sequential sensitivity analysis for liquid food stimuli. J.Food Sci. 51(6):1550-3.

O'Mahony M. 1986. Sensory evaluation of food: statistical methods and procedures. CRC Press.

O'Mahony M, Odbert N. 1985. A Comparison of Sensory Difference Testing Procedures: Sequential Sensitivity Analysis and Aspects of Taste Adaptation. J.Food Sci. 50(4):1055-8.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M. 1992. Understanding Discrimination Tests: A User-friendly Treatment Of Response Bias, Rating And Ranking R-index Tests And Their Relationship To Signal Detection. J.Sens.Stud. 7(1):1-47.

O'Mahony M, Thieme U, Goldstein LR. 1988. The Warm-up Effect as a Means of Increasing the Discriminability of Sensory Difference Tests. J.Food Sci. 53(6):1848-50.

O'Mahony M. 2013. The Tetrad Test: Looking Back, Looking Forward. J.Sens.Stud. 28(4):259-63.

Pellegrino R, Luckett CR, Shinn SE, Mayfield S, Gude K, Rhea A, Seo H. 2015. Effects of background sound on consumers' sensory discriminatory ability among foods. Food Quality and Preference 4371-8.

Peryam DR, Pilgrim FJ, Peterson MS. 1954. Food Acceptance Testing Methodology: A Symposium Sponsored by the Quatermaster Food and Container Institute for the Armed Forces. National Academies 1

Plotto A, Baldwin E, McCollum G, Manthey J, Narciso J, Irey M. 2010. Effect of Liberibacter Infection (Huanglongbing or ?Greening? Disease) of Citrus on Orange Juice Flavor Quality by Sensory Evaluation. J.Food Sci. 75(4):S220-30.

Rousseau B. 2003. Sensory Evaluation | Sensory Difference Testing. In: Benjamin Caballero, editor. Encyclopedia of Food Sciences and Nutrition (Second Edition). Oxford: Academic Press. p 5141-7.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Rousseau B, Ennis JM. 2013. Importance of Correct Instructions in the Tetrad Test. J.Sens.Stud. 28(4):264-9.

Rousseau B, Stroh S, O'Mahony M. 2002. Investigating more powerful discrimination tests with consumers: effects of memory and response bias. Food Quality and Preference 13(1):39-45.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-84.

Schlich P. 1993. Risk tables for discrimination tests. Food Quality and Preference 4(3):141-51.

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Stone H, Bleibaum R, Thomas HA. 2012. Sensory evaluation practices. Academic press.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

Thieme U, O'Mahony M. 1990. Modifications To Sensory Difference Test Protocols: The Warmed Up Paired Comparison, The Single Standard Duo-trio And The A-not A Test Modified For Response Bias. J.Sens.Stud. 5(3):159-76.

Thurstone LL. 1994. A law of comparative judgment. Psychol.Rev. 101(2):266-70.

Ura S. 1960. Pair, triangle and duo-trio test. Reports of statistical application research 7107.

Van Hout D, Hautus MJ, Lee H. 2011. Investigation of test performance over repeated sessions using signal detection theory: comparison of three nonattribute-specified difference tests 2-AFCR, A-not A and 2-AFC. J.Sens.Stud. 26(5):311-21.

Xia Y, Zhang J, Zhang X, Ishii R, Zhong F, O'Mahony M. 2015. Tetrads, triads and pairs: Experiments in self-specification. Food Quality and Preference 40, Part A97-105.

Chapter 3: Beverage Complexity Yields Unpredicted Power Results for 7 Discrimination Test Methods

3.1 Abstract

The power of discrimination tests is crucial in determining sample size and resources needed for testing. Although research has been conducted on the power analysis of several discrimination testing methods, much of the previous research has focused on basic taste solutions, which may not be directly applicable to food and beverage systems. The objective of the current study was to compare the power of seven discrimination tests: Panelist-Articulated-2-Alternative Forced Choice (PA-2-AFC), Triangle, Triangle with Partial Presentation, Duo-trio, Duo-trio with Partial Presentation, 4-category rating methods for R-index measure, and same-different pairwise comparison for R-index measure using four different complex beverage systems.

Sixty-one pre-screened panelists participated in the study. Six product comparisons were performed using tea, tomato juice (3 comparisons), citrus-flavored carbonated soda, and cola beverage systems. The tests conducted in the study were randomized over two testing sessions for each product comparison.

Triangle testing methodologies were found to be overall the most powerful methods across product categories. The PA-2-AFC method was found to be the least powerful across all products. Thurstonian modeling predicts that the PA-2-AFC method would be the most powerful method used in testing contrary to the findings of the current study. Samples used in testing were complex in both basic formulations and in changes made between control and variant samples. Complexity of the samples may have influenced the discriminability by the panelists. Further

research should be conducted to identify the influence of sample complexity on the power of discrimination methodology.

3.2 Introduction

The use of sensory discrimination testing allows researchers to determine if two products are significantly different based on human sensory perception (Lawless and Heymann 1999). In doing so, sensory professionals advise product developers on decisions such as reformulations, production alterations, or quality concerns. As the results of discrimination testing can be used for anything from internal panel assessment to billion dollar brand launches, selecting a powerful sensory discrimination method is highly important. If power is low, a discrimination test cannot reliably detect differences between products (Ennis and Jesionka 2011). This may result in an undetected differences to be noticed by consumers and result in negative perception.

Research conducted in the area of methodology comparison is often performed using model solutions such as sodium chloride in water (Byer and Abrams 1953a; O'Mahony and Odbert 1985; Tedja and others 1994; Mata-Garcia and others 2007), or with foods in which one ingredient alteration has been made during formulation (Stillman 1993; Masuoka and others 1995; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Rousseau and others 1998). While these studies have led to important scientific discoveries into how subjects approach each method, the complex nature of commercial foods and beverages may make the idea of method selection more difficult than previously thought.

Thurstonian modeling predictions are typically based on unidimensional models, which assume that samples differ along one dimensional continuum (Frijters 1980). Multidimensional models have been created (Ennis and Mullen 1985; Ennis and Mullen 1986), which indicate that

a decrease in discriminability may occur when multidimensional samples are utilized in testing. Experimental results must be collected in order to determine how sample dimensionality influences test power as a lack of consistency among predictions and experimental results may occur when multidimensional samples are utilized in testing (Ennis 1998).

Thurstonian model predictions indicate that specific test methods such as the n-AFC methods are more powerful than non-specific methods such as the Triangle test method (Byer and Abrams 1953a; Gridgeman 1970; Frijters and others 1980; Ennis 1990; Ennis 1993; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005). The challenge of utilizing specific test methods such as the n-AFC tests comes from the need to specify an attribute by which subjects are to differentiate the samples. Several ways of circumventing this issue have been developed, which include the use of subjects self-specifying the attribute as well as the use of differentiation based on preference (McClure and Lawless 2010; Xia and others 2015). Utilizing subjects to designate the attribute, with which samples are to be differentiated may allow sensory professionals the ability to utilize specific difference test methods without having to designate a difference for subjects and should be further explored.

The objectives of this study were to 1) compare the power among non-specific difference test methods and Panelist-Articulated-2-Alternative Forced Choice method and 2) compare experimental data to Thurstonian model predictions using complex beverages. It was hypothesized that panelist articulation will allow for the use of a powerful specific discrimination test method over a less powerful non-specific test method when attributes are not easily specified by researchers. Additionally, since Thurstonian model predictions are based on a unidimensional

axis, using multidimensional samples will impact the power of test methods in possibly unpredictable ways depending on sample complexity and correlation of dimensions.

3.3 Materials and Methods

3.3.1 Subjects

Sixty-one subjects were prescreened and recruited to participate in the study. Prescreening procedures included basic taste identification, aroma identification, and discrimination testing (triangle and duo-trio methods). Inclusion in the study required a correct response rate of at least 70% on prescreening samples. In addition to sensory acuity, subjects were selected based on frequent consumption of the sample beverage categories utilized in the study including teas, juices, and carbonated beverages. All subjects were free of food allergies. Subjects received monetary compensation at the completion of the study.

3.3.2 Samples

Three product categories, tea, juice and carbonated beverages were selected for testing to represent a range of products common to the beverage industry. For each product comparison, two confusable samples were chosen to be used in discrimination testing.

For the tea category, one control sample and one variant sample were selected. Tea samples were commercially available lemon-flavored tea products. The control samples were packaged and stored in glass packaging, while the variant samples were packaged and stored in polyethylene terephthalate (PET) bottles. No formulation differences existed between the actual teas placed in the two packaging materials.

Samples used in the vegetable juice category were a commercially available tomato juice product. One control sample and three variant samples were selected for a total of 3 product comparisons. The control sample was the same in each product comparison. Variant samples used had increasing levels of sodium reduction. Juice 1 had the smallest amount of sodium reduction, with increasing reduction in sodium for Juice 2 and Juice 3.

For the carbonated beverage category, one control and one variant sample were selected for each of two different carbonated beverages. One cola-flavored beverage and one citrus-flavored beverage were used in testing. The cola-flavored beverage utilized a different sweetener profile between control and variant samples. An increase in citrus-flavor and citric acid used in the formulation of the citrus-flavored beverage constituted the differences between control and variant samples. Carbonated beverages were served to panelists immediately upon opening of an individual can. No can was held for more than 2 minutes after opening. All samples were stored at 22°C and served at room temperature.

3.3.3 Experimental Procedure

Seven discrimination tests, 2-AFC, Triangle, Triangle with partial presentation, Duo-trio, Duo-trio with partial presentation, 4-category rating method for R-index, and same-different pairwise comparison for R-index, were conducted. A description of the method along with serving orders presented in partial presentation methods can be found in Table 3.1. For each test, panelists performed one replicate test. The order of presentation of samples for each test was randomized across panelists, and all possible orders for 2-AFC, Triangle and Duo-trio tests were equally presented. Triangle with partial presentation was presented as two control samples and one variation. Duo-trio with partial presentation was presented with the control sample as

reference. For the 4-category rating method to determine R-index measure, the order of the presentation was randomized within a panelist as well as across panelists to ensure minimal order bias.

For each product comparison, a total of two sessions were conducted. In one session, four traditional difference tests (Triangle, Triangle with partial presentation, Duo-trio, and Duo-trio with partial presentation) were conducted with five min break in between each test. In the other session, the two methods to obtain R-index and a PA-2-AFC were conducted with a five min break in between each test. The two sessions were randomized among panelists, and the tests within a session were randomized, so as to minimize order bias across different tests. For a total of six product comparisons (three product types x 1-3 different levels compared to the control product), there were a total of 12 sessions that were conducted per panelist.

For PA-2-AFC tests, the procedure was as follows. Prior to the actual 2-AFC test, an articulation procedure was conducted. For this, panelists were given a set of the two samples to be tested, and tasted the samples alternately from each, until they were able to articulate the difference. The tasting of alternate samples was performed as rapidly as possible to allow differences to be perceived, so as to facilitate an efficient signal search. Once the difference was identified by the panelist, s/he articulated what the difference was and which sample had more of that attribute on the warm-up ballot. This information was transferred to the standard 2-AFC ballot for actual testing of the samples.

For the R-index measure by the 4-category rating method, a randomized complete block design was used. One sample was designated as the noise and the other as the signal. Panelists performed a warm-up with the noise sample, as to familiarize themselves with the characteristics of the noise sample. After the panelist was able to become familiar with the noise sample, s/he

began the experiment. One replicate of the noise and the signal was given as a complete set per panelist. Panelists were instructed to take the whole sample into the mouth, swirl it around for 2–3 s, expectorate, and complete the task given during the experiment. Samples were presented monadically. Panelists were asked to rate the sample on the 4-category rating scale, where the categories are "signal sure", "signal unsure", "noise sure", and "noise unsure" (Figure 3.4). Panelists were instructed to rinse after every sample to reduce adaptation and to be consistent with other discrimination tests being compared.

For the R-index measure by the same-different pairwise comparison (Figure 3.5), a randomized complete block design was used. Two samples were presented simultaneously, and panelists were asked if the two samples are same or different with the sureness judgment as described above for the 4-category rating method. One replicate of the noise and the signal were given as a complete set per panelist. The tasting procedure was the same as the R-index by the 4-category rating method.

All tests were conducted in booths where the temperature was set approximately at 22°C and relative humidity at 33%. To eliminate possible carryover effects, a rinse protocol was developed for all product types. The rise protocol was a three step rinse process beginning with carbonated water, followed by warm water (43-49°C), and ending with room temperature water. Panelists began the test by rinsing the mouth with the rinse protocol. They then performed each test with interstimulus rinsing. No color differences were observed between the samples, so all testing was performed under incandescent lighting. The order of test sessions was counterbalanced over panelists, so as to minimize order bias across different tests. Compusense® *five Plus* (Version 4.6: Guelph ON, Canada) program was used for data collection.

3.3.4 Data Analysis

For Triangle, Duo-trio, PA-2-AFC tests, binomial test was conducted to determine if significant difference exists between the two samples. Power was calculated for each test using the number of correct responses across all panelists' data. Power calculations were made using a sample size of 61, significance level of 0.05, and the calculated delta value (d'-value) for each test and product combination. For the PA-2-AFC method utilizing the citrus carbonated beverages, a sample size of 60 was utilized in power analysis as one subject indicated they were unable to determine a difference between the samples during the articulation procedures. Data analysis was performed using IFPrograms™ version 8.11 (The Institute for Perception, USA).

For the R-index measure, Table 3.2 was used to tabulate the number of responses elicited for each category of responses for signal and noise samples. With the response matrix, R-index was calculated using the equation found in Table 3.2 (b).

To determine if significant differences existed between d' values from different test methods for the same product, Chi-Square analysis was conducted using IFPrograms™ version 8.11 (The Institute for Perception, USA). For methods which resulted in a d' value of 0, the variance of d' is essentially infinite (Ishii and others 2014). These values were omitted from Chi-Square analysis. Chi-Square results can be found in Table 3.3. Products in which one of the methods resulted in a d' value of 0 were compared to the method that resulted in the highest d' value for the same product. To do so, the highest d' value for the product was compared to 0 by obtaining the Z-value using the following equation (Bi and others 1997):

$$Z = \frac{d'}{\sqrt{S_{d'}^2}}$$

d' comparison results can be found in Table 3.4.

3.4 Results and Discussion

When analyzing the results of the discrimination testing as a whole, no difference test was observed to be universally most powerful across all product categories (Tables 3.5-3.7). The PA-2-AFC method was found to have the lowest power across all products contrary to expected results. Comparing individual product categories allows for further exploration of the resulting data.

Within the carbonated beverage product category (Table 3.5), the most powerful testing method of the standard tests used across all products was the Triangle test with partial presentation. The Paired Comparison R-Index measure resulted in the greatest significance for the diet citrus carbonated beverage (p-value 0.001), but does not appear to be stable within the carbonated beverage category itself as the method resulted in the least significant results (p-value 0.357) for the diet cola carbonated beverage.

For the Juice product category (Table 3.6), Juice 2 appeared to be an anomaly for many testing methods. Based on the changes made to the juice product, an increasing reduction in sodium with the least reduction in Juice 1 to the greatest reduction in Juice 3, it is expected that a general increase in discriminability between the control product and variation tested. Actual results indicate a decrease in discrimination, based on p-value, between Juice 1 and Juice 2 for PA-2-AFC (0.50 to 0.695), Paired-Comparison R-Index (0.005 to 1.0), Triangle with partial presentation (0.19 to 0.28), and Triangle with balanced presentation (0.08 to 0.37) and the same discriminability for Duo-trio with balanced presentation (0.30) and R-Index measure (0.20).

For the Tea product (Table 3.7), the most powerful testing method was determined to be the Triangle test with partial presentation (power = 99.5%) followed by the Paired Comparison

R-Index measure and Triangle with balanced presentation. The PA-2-AFC and Duo-Trio with balanced presentation were the least powerful methods of this category.

Across all product categories, the PA-2-AFC test was one of the least powerful methods used. The overall low power for the PA-2-AFC does not support our hypothesis that the specific difference tests (i.e., 2-AFC) would be more powerful than the general difference tests (i.e., duo-trio), based on the literature, which has shown that a specific difference test, such as the 3-AFC, is statistically more powerful than a non-specific difference test, such as the Triangle test (Byer and Abrams 1953a; Gridgeman 1970; Frijters and others 1980; Ennis 1990; Ennis 1993; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; O'Mahony 1995; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005).

The increase in the power of specific difference test methods, originally described as the paradox of discriminatory non-discriminators (Gridgeman 1970), has been explained by the change in discriminal processes from a comparison of distances strategy used in non-specific tests to a skimming strategy used in specific difference tests as predicted by Thurstonian modeling (Frijters 1979; O'Mahony 1995). When a subject is asked to identify the odd sample as in the triangle test, subjects compare the distance in perceived intensity of the samples and will choose the most different sample in terms of perceived intensity. When a subject is asked to identify the sample with the greatest perception of a single attribute, i.e. sweetness, as in the AFC tests, subjects will choose the sample with the greatest perceived intensity of the identified attribute.

There are several theories to explain the unexpected results found in the current testing. Sample complexity may have influenced the discriminability of the samples by subjects (Ennis and Mullen 1985; Ennis and Mullen 1986; Mullen and Ennis 1987). When comparing d’ values

of different methods for the same product, significant differences ($p<0.05$) between d' values were observed for all products except Juice 1. When comparing multiple methods using Chi-Square analysis, the Diet Cola and Juice 3 products had significant differences between d' values. Three of the six products included in testing resulted in a d' of 0 for at least one of the methods tested. The methods with a d' of 0 could not be included in Chi-Square analysis. These d' values were compared to the highest d' value of a method using the same product to determine if significant differences exist between the highest d' value and 0 for each product. The 3 products which were compared in Table 3.4 all resulted in significant differences between d' values.

Products tested exhibit complexity in both basic formulation as well as in the formula changes made between control and variant products. While much of the research on determination of power differences between specific and non-specific difference tests is performed using basic taste solutions and food products with only one ingredient alteration (Byer and Abrams 1953b; O'Mahony and Odbert 1985; Tedja and others 1994; Rousseau and others 1998; Rousseau and others 2002; Lau and others 2004; Braun and others 2004; Liggett and Delwiche 2005; Angulo and others 2007; McClure and Lawless 2010), when applying these methods to complex food products, unidimensional Thurstonian modeling predictions may no longer be applicable to discrimination testing methodology (Ennis 1998). Even in product categories where only one ingredient or packaging change was made, the resulting changes to the product may be complex, as in the tea product category where a change from packaging in glass to PET may alter a great number of product attributes. A packaging change may generate complex changes between samples due to imparting of flavor from packaging material, or scalping of flavor compounds common to packaging in PET (Sajilata and others 2007). Another

theory which may explain poor performance of 2-AFC test in the current study is the need for panelists to articulate the nature of the difference for each sample. As samples tested were complex products, the overall differences between each sample may have been greater than any one attribute that the panelist may articulate. A combined total of the differences in this case may allow for greater panel performance by a non-specific difference test such as the triangle test.

Two methods were used to determine the R-Index measures, R-Index measure by the Paired Comparison method (PC R-Index) and R-Index measure by the 4-category rating method (R-Index). R-index indicates the probability of discriminating between samples when presented in a paired comparison presentation. The PC R-Index measures resulted in high discrimination based on p-value for many products tested including diet citrus carbonated beverage (p-value = 0.001), Juice 1 (p-value =0.005), Juice 3 (p-value = 0.001), and Tea (p-value = 0.002). Within these product categories the method was within the two most significant resulting test methods. However, the PC R-Index measure resulted in low discrimination based on p-value for two products tested, which were diet cola carbonated beverage (p-value = 0.357) and Juice 2 (p-value = 1.000).

The R-Index measure by the 4-category rating method resulted in middle to low range discrimination between samples across all product categories. While other discrimination testing methods were able to identify the increasing level of sodium reduction with increasing levels of discrimination, the R-Index measure by 4-category rating method resulted in the same level of discrimination across all juice products. As this product category consisted of products with an increasing level of reduction of a single ingredient within a complex flavor matrix, the R-Index measure by 4-category rating method may not be suitable for use in testing subtle changes in formulation within complex foods and beverages based on results observed in the current study.

3.5 Conclusions

The current study has found discrimination testing methodology power to be influenced by sample product category and may be impacted by the complexity of samples being tested. There are several possible future directions in research leading from the findings from the present study. Results have shown that applying methodology based on basic taste solutions and food products with only one ingredient change may not directly relate to complex food systems.

Complexity arises in two forms in food and beverage products. Firstly, a food product may have a complex flavor profile due to the multidimensional formulation of the product. An example from the current study of a complex matrix food product may be the juice product category, where many flavors and tastes make up the profile of the product. Secondly, a food may have a complex formulation change between control and variant samples. An example of this type of complexity found in the current study may be the tea category where flavor scalping between packaging can occur. Both forms of sample complexity and multidimensionality may impact discrimination testing results and are factors, for which basic Thurstonian modeling does not account.

Further research should be conducted to explore how complex samples with several formulation changes or varying ingredient interactions may impact the power of sensory discrimination testing methods. Possible studies may look into the impact of concurrent changes made to a beverage or food product in relation to the expected impact on discrimination testing methodology power. As seen in the Juice product category, changing the concentration of one ingredient in a complex product may have varying impact on sensory discrimination.

The human as a sensory instrument is impacted by a variety of factors. From the environment that testing takes place, to the vessel which samples are served, many factors can be controlled and manipulated by the scientist. What we have little control over, and which has a great impact on results, is what occurs within the nervous system to allow subjects to identify differences. The overlapping neural pathways which enable our ability to perceive lead to unpredictability in response when complex stimuli are utilized. Nevertheless, this unpredictability should be explored to strengthen the tools available to sensory scientists.

3.6 References

Angulo O, Lee H, O'Mahony M. 2007. Sensory difference tests: Overdispersion and warm-up. Food Quality and Preference 18(2):190-5.

Bi J, Ennis DM, O'Mahony M. 1997. How to estimate and use the variance of d' from difference tests. J.Sens.Stud. 12(2):87-104.

Braun V, Rogeaux M, Schneid N, O'Mahony M, Rousseau B. 2004. Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. Food Quality and Preference 15(6):501-7.

Byer AJ, Abrams D. 1953a. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Byer AJ, Abrams D. 1953b. A comparison of the triangular and two-sample taste-test methods. Wallerstein Lab Commun 16((54)):253-60.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM. 1998. Thurstonian Scaling for Difference Tests. IFPress 1(3):2.

Ennis DM, Mullen K. 1986. A multivariate model for discrimination methods. J.Math.Psychol. 30(2):206-19.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Ennis DM, Mullen K. 1985. The effect of dimensionality on results from the triangular method. Chemical Senses 10(4):605-8.

Ennis JM, Jesionka V. 2011. The Power Of Sensory Discrimination Methods Revisited. J.Sens.Stud. 26(5):371-82.

Frijters JER. 1980. Three-stimulus procedures in olfactory psychophysics: An experimental comparison of Thurstone-Ura and three-alternative forced-choice models of signal detection theory. Percept.Psychophys. 28(5):390-7.

Frijters JER. 1979. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER, Kooistra A, Vereijken PFG. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Percept.Psychophys. 27(2):176-8.

Gridgeman NT. 1970. A Reexamination of the Two-Stage Triangle Test for the Perception of Sensory Differencesa. J.Food Sci. 35(1):87-91.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Qual.Preference 31(1):49-55.

Lau S, O'mahony M, Rousseau B. 2004. Are three-sample tasks less sensitive than two-sample tasks? Memory effects in the testing of taste discrimination. Percept.Psychophys. 66(3):464-74.

Lawless HT, Heymann H. 1999. Discrimination Testing. In: R. Bloom, editor. Sensory Evaluation of Food. Springer US. p 116.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Masuoka S, Hatjopoulos D, O'Mahony M. 1995. Beer Bitterness Detection: Testing Thurstonian And Sequential Sensitivity Analysis Models For Triad And Tetrad Methods. J.Sens.Stud. 10(3):295-306.

Mata-Garcia M, Angulo O, O'Mahony M. 2007. On Warm-up. J.Sens.Stud. 22(2):187-93.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

Mullen K, Ennis DM. 1987. Mathematical formulation of multivariate euclidean models for discrimination methods. Psychometrika 52(2):235-49.

O'Mahony M, Odbert N. 1985. A Comparison of Sensory Difference Testing Procedures: Sequential Sensitivity Analysis and Aspects of Taste Adaptation. J.Food Sci. 50(4):1055-8.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M. 1992. Understanding Discrimination Tests: A User-friendly Treatment Of Response Bias, Rating And Ranking R-index Tests And Their Relationship To Signal Detection. J.Sens.Stud. 7(1):1-47.

O'Mahony M. 1995. Who told you the triangle test was simple? Food Quality and Preference 6(4):227-38.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivit of the same-different test: Comparison with triangle and duo-trio methods. J.Sens.Stud. 13(2):149-73.

Rousseau B, Stroh S, O'Mahony M. 2002. Investigating more powerful discrimination tests with consumers: effects of memory and response bias. Food Quality and Preference 13(1):39-45.

Sajilata MG, Savitha K, Singhal RS, Kanetkar VR. 2007. Scalping of Flavors in Packaged Foods. Comprehensive Reviews in Food Science and Food Safety 6(1):17-35.

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

Xia Y, Zhang J, Zhang X, Ishii R, Zhong F, O'Mahony M. 2015. Tetrads, triads and pairs: Experiments in self-specification. Food Quality and Preference 40, Part A97-105.

3.7 Tables and Figures

Table 3.1. Description of sensory discrimination testing methods utilized in testing.

Test Method	Task	Sample Presentation Orders	Chance Probability
Panelist-Articulated-2-AFC	Identify the difference between sample A and sample B Identify the more "_____" sample	AB, BA	1/2
Triangle	Identify the odd sample	AAB, ABA, BAA, BBA, BAB, ABB	1/3
Triangle with Partial Presentation	Identify the odd sample	AAB, ABA, BAA	1/3
Duo-trio	Identify the sample that is the same as the reference	RA:AB, RA:BA, RB:BA, RB:AB	1/2
Duo-trio with Partial Presentation	Identify the sample that is the same as the reference	RA:AB, RA:BA	1/2
4-Category Rating Method for R-Index	Identify if the sample is the same as the noise (with sureness)	Noise A:AB, Noise A:BA	
Same-Different Pairwise Comparison for R-Index	Identify if the samples are the same or different (with sureness)	AA, BB, AB, BA	

Table 3.2. (a) Response matrix of R-Index using signal detection rating method

	SS	S?	N?	NS
Signal	a	b	c	d
Noise	e	f	g	h

Table 3.2. (b) Equation for the computation of R-Index. Letters found in the equation correspond to those in the response matrix (Table 3.2 (a))

$$\mathrm{R-index} = \frac{a(f+g+h) + b(g+h) + ch + \frac{1}{2}(ae + bf + cg + dh)}{(a+b+c+d)(e+f+g+h)}$$

Table 3.3. Comparison of discrimination test methods d' values for the same product. Methods with d' values of 0 were not included in analysis. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 indicates the variance of d'. Bolded p-values are below 5% indicating significant differences between d' values for the same product across methods listed.

Product	Test	d'	σ^2 of d'	chi-square	critical	p-value
Diet Cola	Duo-Trio Partial††	1.36	0.135	18.402	9.488	**0.001**
	Duo-Trio Regular	1.45	0.129			
	PA-2-AFC†	0.15	0.052			
	Triangle Partial	1.51	0.111			
	Triangle Regular	0.83	0.210			
Diet Citrus	Duo-Trio Partial	0.30	1.401	0.174	3.841	0.676
	Triangle Partial	0.83	0.210			
Juice 1	Duo-Trio Partial	0.94	0.198	7.863	9.488	0.097
	Duo-Trio Regular	0.69	0.316			
	PA-2-AFC	0.03	0.052			
	Triangle Partial	0.83	0.210			
	Triangle Regular	1.05	0.154			
Juice 2	Duo-Trio Partial	1.05	0.172	0.592	7.815	0.898
	Duo-Trio Regular	0.69	0.316			
	Triangle Partial	0.70	0.271			
	Triangle Regular	0.55	0.406			
Juice 3	Duo-Trio Partial	1.26	0.143	14.028	7.815	**0.003**
	Duo-Trio Regular	0.53	0.496			
	PA-2-AFC	0.15	0.052			
	Triangle Regular	1.51	0.111			
Tea	Duo-Trio Partial	1.26	0.143	4.650	7.815	0.199
	Duo-Trio Regular	0.30	1.401			
	Triangle Partial	2.09	0.102			
	Triangle Regular	1.42	0.115			

†PA-2-AFC: Panelist-Articulated-2-Alternative Forced Choice

††Partial: Indicates only half of possible serving orders were presented during testing

Table 3.4. Methods with a d' = 0 were compared to the highest d' value from a method using the same products to determine if significant differences existed between d' values from difference methods for the same products. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 indicates the variance of d'. Bolded p-values are below 5% indicating significant differences between d' values for the same product across methods listed.

Product	Test	d'	σ^2 of d'	z-value	p-value
Diet Citrus	PA-2-AFC†	0	N/A		
	Triangle Partial†	0.83	0.210	1.810	**0.035**
Juice 2	PA-2-AFC	0	N/A		
	Duo-Trio Partial	1.05	0.172	2.530	**0.006**
Tea	PA-2-AFC	0	N/A		
	Triangle Partial	2.09	0.102	6.558	**0.000**

†PA-2-AFC: Panelist-Articulated-2-Alternative Forced Choice

††Partial: Indicates only half of possible serving orders were presented during testing

Table 3.5. P-value and power for carbonated beverage products across methods. P value listed indicates the significance at which the test method can declare the sample pair different. Bolded p-values are below 5% indicating the test method declared samples significantly different at alpha=0.05. Power is the probability of declaring a difference when one exists between samples.

Product	Test	P-value	Power
Diet Cola	Duo-Trio Partial††	**0.020**	0.662
	Duo-Trio Balanced	**0.010**	0.748
	PA-2-AFC†	0.304	0.127
	R-Index	**0.050**	N/A
	PC-R-Index	0.357	N/A
	Triangle Partial	**0.004**	0.876
	Triangle Balanced	0.194	0.253
Diet Citrus	Duo-Trio Partial	0.500	0.047
	Duo-Trio Balanced	0.980	0.036
	PA-2-AFC	0.963	0.046
	R-Index	>0.4	N/A
	PC-R-Index	**0.001**	N/A
	Triangle Partial	0.194	0.253
	Triangle Balanced	0.905	0.049

†PA 2-AFC: Panelist-Articulated 2-Alternative Forced Choice

††Partial: Indicates only two weaker and one stronger serving orders were presented during testing

Table 3.6. P-value and power for juice products across methods. P value listed indicates the significance at which the test method can declare the sample pair different. Bolded p-values are below 5% indicating the test method declared samples significantly different at alpha=0.05. Power is the probability of declaring a difference when one exists between samples.

Product	Test	P-value	Power
Juice 1	Duo-Trio Partial††	0.153	0.260
	Duo-Trio Balanced	0.304	0.125
	PA-2-AFC†	0.500	0.036
	R-Index	0.200	N/A
	PC-R-Index†††	**0.005**	N/A
	Triangle Partial	0.194	0.253
	Triangle Balanced	0.082	0.444
Juice 2	Duo-Trio Partial	0.100	0.350
	Duo-Trio Balanced	0.304	0.125
	PA-2-AFC	0.695	0.036
	R-Index	0.200	N/A
	PC-R-Index	1.000	N/A
	Triangle Partial	0.275	0.175
	Triangle Balanced	0.371	0.115
Juice 3	Duo-Trio Partial	**0.036**	0.558
	Duo-Trio Balanced	0.399	0.080
	PA-2-AFC	0.304	0.127
	R-Index	0.200	N/A
	PC-R-Index	**0.001**	N/A
	Triangle Partial	0.591	0.049
	Triangle Balanced	**0.004**	0.876

†PA 2-AFC: Panelist-Articulated 2-Alternative Forced Choice

††Partial: Indicates only two weaker and one stronger serving orders were presented during testing

†††PC: Paired Comparison

Table 3.7. P-value and power for tea product across all testing methods. P value listed indicates the significance at which the test method can declare the sample pair different. Bolded p-values are below 5% indicating the test method declared samples significantly different at alpha=0.05. Power is the probability of declaring a difference when one exists between samples.

Product	Test	P-value	Power
Tea	Duo-Trio Partial††	**0.036**	0.558
	Duo-Trio Balanced	0.500	0.047
	PA-2-AFC†	0.900	0.036
	R-Index	**0.020**	N/A
	PC-R-Index†††	**0.002**	N/A
	Triangle Partial	**0.000**	0.995
	Triangle Balanced	**0.007**	0.720

†PA 2-AFC: Panelist-Articulated 2-Alternative Forced Choice

††Partial: Indicates only two weaker and one stronger serving orders were presented during testing

†††PC: Paired Comparison

Table 3.8. R-Index measures and p-values for Paired Comparison R-Index and R-Index by 4 category rating method across all product categories tested. Bolded p-values are below 5% indicating the test method was able to declare samples significantly different at alpha=0.05.

Product	Test	R-Index	P-value
Diet Cola	R-Index	0.63	**0.050**
	PC-R-Index†	0.69	0.357
Diet Citrus	R-Index	0.50	>0.4
	PC-R-Index	0.84	**0.001**
Juice 1	R-Index	0.60	0.200
	PC-R-Index	0.83	**0.005**
Juice 2	R-Index	0.60	0.200
	PC-R-Index	0.50	1.000
Juice 3	R-Index	0.60	0.200
	PC-R-Index	0.91	**0.001**
Tea	R-Index	0.66	**0.020**
	PC-R-Index	0.84	**0.002**

†PC: Paired Comparison

Chapter 4: Warm-up Effect in Panelist-Articulated-2-Alternative Forced Choice Test

4.1 Abstract

Panelist performance in discrimination tests has shown to increase when warm-up samples are provided prior to the actual test. Samples are used prior to the actual test for the attribute articulation process of a Panelist-Articulated-2-Alternative Forced Choice (PA-2-AFC) procedure; however, it is yet unknown if the pretest articulation phase adds to the power of this testing method as with the warm-up. The goal of the study was to determine if a "warm-up" effect was displayed in the PA-2-AFC test resulting in greater power compared to the Researcher-Designated-2-AFC (RD-2-AFC) test, typically used in sensory discrimination testing.

A RD-2-AFC test, with and without warm-up samples, and a PA-2-AFC test were performed by 61 pre-screened panelists. A reduced calorie, citrus-flavored, carbonated beverage was used in the testing procedures. The two test samples differed in the levels of citric acid and citrus flavor; thus, during RD-2-AFC testing, panelists were asked to identify which sample was more sour. For PA-2-AFC testing, panelists individually articulated the nature and direction of the difference between the two samples through a pre-testing articulation procedure. The articulated difference was, then, used in standard 2-AFC test procedure.

A warm-up effect was observed between the standard RD-2-AFC without warm-up samples and RD-2-AFC with warm-up samples, with a significant increase in power with the addition of warm-up samples. However, the PA-2-AFC method was shown to be the least powerful method.

The increase in power with the addition of warm-up samples for the RD-2-AFC procedure supports literature findings on the benefit of providing warm-up samples. No warm-up

effect can be attributed to the PA-2-AFC method evidenced by the overall low power observed, which may be attributed to sample complexity. Future research should be conducted to determine how sample complexity impacts panelists' ability to articulate differences in the pretest articulation process.

Keywords: 2-AFC, Warm-up, Discrimination testing, Panelist Articulated

4.2 Introduction

Identification of powerful sensory discrimination testing methods has been an area of focus for sensory scientists for several decades. Emphasis has been placed on methodology improvement, as methods with more power, or the reliability at which a test method can detect differences between products (Ennis and Jesionka 2011), can influence the amount of resources needed to perform a discrimination test. Once such improvement has been a focus on what question subjects are asked to answer. Simple changes in the instructions given to a subject during a sensory discrimination test may result in a reduction of resources needed to perform a discrimination test in the form of both time and physical materials. For instance, a more powerful test method would require a smaller sample size (Rousseau 2003), thus reducing the resources needed to perform the test. Utilizing a more powerful test method is preferred, as a low power test method may result in an important difference between samples being missed (Bi and Ennis 1999). This missed difference would lead a sensory scientist to interpret the samples as being not significant different, when they could have been declared different with a more powerful method.

One way which has been proposed to increase subject performance, thus increase power of a test method, is through the utilization of warm-up samples prior to actual testing. First

identified in the field of experimental psychology (Heron 1928), a warm-up effect is the period of time in which subject's performance rapidly improves as they become acquainted with the task for which they are being asked to perform. In order to induce a warm-up effect, subjects are asked to rapidly taste pairs of samples to be tested up to 10 times until they become familiar with the difference between the two samples. During this period, subjects actively search for differences and acquaint themselves with the sample matrix. Introduction of warm-up samples has been shown to increase subject performance by inducing a warm-up effect (O'Mahony and others 1988; Thieme and O'Mahony 1990; Dacremont and others 2000; Mata-Garcia and others 2007; Angulo and others 2007).

Another way in which to improve performance on sensory discrimination testing is the selection of a specified, or directional method, over a non-specified, or non-directional method. Based on Thurstonian modeling predictions, changes in the decision rule account for observed enhancement in the proportion of correct responses when a specified method is utilized. For example, when comparing the 3-AFC method to the triangle test method, methods that differ only by the instructions given to the subject, a significantly greater proportion of discriminators is theorized and observed for the 3-AFC method over the triangle test method. (Byer and Abrams 1953; Gridgeman 1970; Frijters and others 1980; Ennis 1990; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005). By focusing on a specific attribute, subjects can use a decision rule which leads to a greater proportion of correct responses, a change which accounts for the increased performance on specified discrimination testing methods.

The inherent challenge for a sensory scientist in deciding to use a specified difference test is the need for an attribute to be selected by which subjects will differentiate samples. Though

not commonly utilized, utilizing warm-up samples to generate an attribute has been suggested as a mode of overcoming this challenge (Thieme and O'Mahony 1990; McClure and Lawless 2010). During a warm-up period subjects may be asked to specify the nature of the difference between the samples. By doing so, a researcher is able to circumvent the challenge of articulating a difference by which subjects are to differentiate samples. Additionally, utilizing this procedure allows each subject to use their own terminology with which to differentiate samples. By allowing subjects to utilize their own language, subjects may select an attribute with which they believe the samples to most greatly differ.

What is not fully understood about utilizing a pre-articulation procedure to designate the attribute of difference for use in a specified discrimination test method is how pre-articulation samples may act as warm-up samples, which may affect test performance compared to when a researcher provides a term for the subjects. Thus, the objectives of the study were to: (1) asses if a warm-up effect can be attributed to the pretest samples utilized in the articulation process of the Panelist-Articulated-2-AFC test method and (2) compare the power of the Panelist-Articulated-2-AFC test method to the Researcher-Designated-2-AFC test method typically used in sensory discrimination testing. It was hypothesized that when samples used in discrimination testing are multidimensional with complex formulation changes, panelists will have an increased proportion of correct responses when allowed to articulate the nature of the difference between samples than when the researcher designates the nature of the difference.

4.3 Materials and Methods

4.3.1 Subjects

Sixty-one prescreened subjects (49 female, 12 male, age range 18-55 years) were recruited from the University of Illinois at Urbana-Champaign campus. Subjects were prescreened based on basic taste identification, aroma identification, and performance on discrimination testing using Duo-Trio and Triangle methods, consumption rate of carbonated beverages, absence of food allergies, and availability during testing schedule. Subjects were compensated monetarily for participation in the study.

In the basic taste identification screening subjects were asked to perform two tasks. In the first task, subjects were presented with three coded samples each of which composed of a basic taste compound in water (Table 4.1). Subjects were asked to taste and identify which basic taste was present in each solution. In the second task, subjects were asked to rank basic taste intensities. To do so, subjects were presented with three sets of three solutions (Table 4.2). Within each set, one basic taste was identified to the subjects and three samples of increasing concentration were presented in a random order.

For the aroma identification screening, subjects were presented with three samples and asked to match the aroma present to those on a list provided to them. Samples consisted of cotton soaked with added commercially available aroma extracts placed in 162 mL sample cups (Table 4.3).

Discrimination testing performance screening consisted of two discrimination test method, the Triangle test and the Duo-Trio test. Samples used in the screening process consisted of a commercially available juice product, Motts for Tots Apple (Dr Pepper Snapple Group Inc., Plano, TX) with added sugar. Control samples consisted of the bottled product and variant

samples consisted of 3% added sugar in the juice. In the Triangle test, subjects were presented with three samples, two of which were the same and one different. Subjects were asked to identify the odd sample and why they thought it was different. In the Duo-Trio test, subjects were presented with the same juice samples. In this test, three samples were presented one labeled as a reference and two coded samples. Subjects were informed that one of the coded samples was the same as the reference and one was different. Subjects were asked to identify which sample was different and identify they thought the sample was different. In order to be included in the study, subjects must have obtained 70% or more correct answers in the prescreening procedures.

Those subjects who qualified for the study were also asked to complete a brief demographic questionnaire which included frequency of consumption for common food and beverage items. Subjects must state that they were consuming carbonated beverages at least once per month to be included in the study. A total of 119 subjects participated in the screening procedures, of which 61 obtained 70% or more correct answers in the prescreening procedures and were included in the final study.

4.3.2 Samples

A commercial citrus flavored, low calorie, carbonated soft drink was utilized for testing. Samples were produced in a commercial pilot plant facility and packaged in 355 mL aluminum cans prior to testing. At the beginning of each discrimination test, cans were opened and 44 mL of each beverage was poured into 60 mL sample cups labeled with random 3-digit codes. Samples were served lidded and immediately served to subjects to preserve carbonation. No can was held for more than 2 minutes after opening. All samples were served at room temperature

(22°C). One control and one variant sample were used in all three test conditions. As the variant sample formulation included increased citric acid and citrus flavor as compared to the control sample, during Researcher-Designated procedure, subjects were asked to differentiate samples based on sourness.

4.3.3 Experimental Procedure

Protocol for Discrimination Testing Methods

Three test conditions were used: Panelist-Articulated-2-AFC, Researcher-Designated-2-AFC with warm-up samples, and Researcher-Designated-2-AFC without warm-up samples. Details of each test method are described below.

Panelist-Articulated 2-AFC (PA-2-AFC) Procedure

In the PA-2-AFC test, subjects began with a pretest ballot and two samples identical to those presented in the actual 2-AFC test. Subjects were instructed to taste each sample and describe the nature of the difference in their own words and indicate if a difference between the samples was identified. Further, they were asked to indicate the direction of difference by stating which sample had more of the specified attribute. If a subject was unable to articulate a difference with the first set of samples they were given an additional set of samples and instructed to repeat the process. This procedure continued until a difference was articulated or until 10 sets of samples were tasted by the subject, whichever came first.

The flexibility in the number of samples tasted by each subject was intended to allow subjects to familiarize themselves with the differences between samples and the option of

declaring the inability to determine a difference between samples. A majority of subjects were able to articulate a difference within two sample pairs. Only one subject indicated that they were not able to articulate the difference between samples after tasting 10 sample pairs. When an attribute was identified and direction of the difference specified, i.e. sample A is sweeter than sample B, the individual subject's term was transferred to a test ballot for completion of the 2-AFC test.

On the test ballot, subjects were instructed to identify the sample which had more of the attribute specified during the articulation process. All possible serving orders were presented equally across subjects to limit sample order bias.

Researcher-Designated-2-AFC (RD-2-AFC) without Warm-up Procedure

Subjects were asked to identify which sample was more sour than the other. No additional samples were presented prior to testing. Sample order was randomized across all subjects to limit sample order bias.

Researcher-Designated-2-AFC (RD-2-AFC) with Warm-up Procedure

Prior to the actual 2-AFC test, a warm-up procedure was conducted. Similar to the articulation process used in the PA-2-AFC method, during the warm-up procedure, subjects were given a set of the two samples identical to those used in actual data collection and were instructed to taste the samples alternately until they felt familiar with the differences between the two samples. Subjects were permitted additional samples if requested during the warm-up procedure. Once a subject felt comfortable with the warm-up samples, they completed the RD-2-

AFC method as described in the RD-2-AFC without warm-up procedure. Sample order was randomized across all subjects to limit sample order bias.

Experimental Design

A total of 2 testing sessions were performed by each subject with the first test session utilizing the PA-2-AFC method and the second session utilizing the RD-2-AFC method with and without warm-up procedures.

The Panelist-Articulated method was performed prior to the Researcher-Designated methods for all subjects. By doing so, subjects were able to articulate the attribute of difference for the samples without prior influence of attributes designated by the researcher. During this session subjects also performed two non-specific discrimination tests to limit the impact of incidental training (McBride and Laing 1979; Frijters and others 1982) caused by the necessary lack of test method randomization used in the current study. The two sessions were spaced at least 24 hours. To ensure that a warm-up effect could not be attributed to samples used in a previous test method, the RD-2-AFC test without warm-up procedure was performed prior to the RD-2-AFC test with warm-up procedure for all subjects.

All tests were conducted in isolated sensory booths where the temperature was set approximately at 22°C and relative humidity at 33%. To eliminate possible carryover effects, a rinse protocol was developed. The rinse protocol was a three step rinse process beginning with carbonated water, followed by warm water (approximately 43°C), and ending with room temperature water (22°C). Subjects began each test by rinsing the mouth with the rinse protocol. They, then, performed each test with interstimulus rinsing. As no color differences were observed between the samples, all testing was performed under incandescent lighting.

Compusense® *five Plus* (Version 4.6: Guelph ON, Canada) program was used for data collection.

4.3.4 Data Analysis

Data analysis was performed using IFPrograms™ version 8.11 (The Institute for Perception, USA). In order to determine if differences in d' values for the three discrimination methods existed, d', an estimation of delta, as well as its variance were calculated. Additionally, the software was utilized in calculation of the level of power of each method. Power calculations were performed using an alpha level of 0.05, n=61 (n=60 for PA-2-AFC method), and the d' value obtained for each method.

4.4 Results and Discussion

The d' values for each method can be found in Table 4.4. The PA-2-AFC method resulted in a percentage of correct responses below the chance level probability. As such, the estimated d' value for this method was 0 and the variance of d' infinite (Ishii and others 2014). Thus, the d' values for this method were not used in the calculation to determine differences among methods. The resulting d' values from the RD-2-AFC method with and without warm-up procedure, 0.70 and 0.20 respectively, were not significantly different ($\alpha = 0.05$). While the d' values for each method were not significantly different, large differences were present in test power.

The RD-2-AFC without warm-up resulted in a non-significant difference between the two samples tested (p=0.221), while the RD-2-AFC with warm-up resulted in a significant difference between the two samples (p=0.002). As the percentage of correct responses for the

PA-2-AFC method was below the chance probability of 0.5, no significant difference was observed between the two samples tested when using this method. With the addition of warm-up samples in the Researcher-Designated methods, a warm-up effect was observed as seen in the increased power (Figure 4.1) of the RD-2-AFC with warm-up procedure. At α=0.05, n=61, and utilizing each method's corresponding d' value, the warm-up procedure resulted in an increase in power from 0.179 with no warm-up samples to 0.896 with warm-up samples. The observed "warm-up" effect indicates that providing warm-up samples can greatly increase panelists' discrimination. The observed increase in panel performance supports literature findings on the benefit of providing warm-up samples prior to testing procedures using Researcher-Designated methods (O'Mahony and others 1988; Angulo and others 2007).

Results from the PA-2-AFC method indicate that the samples used in the articulation process prior to testing do not provide a warm-up effect to increase performance. Additionally, the articulation process itself caused a decrease in panel performance, as evidenced by the observed decrease in correct responses. This decrease in panel performance may be attributed to subjects identifying perceptual noise variations as opposed to sample differences caused by formulation changes.

The d' for the samples used in the study was below 1 for all testing methods used. As this is a relatively small degree of difference, the perceptual differences caused by formulation changes may be at a level close to that caused by the subjects' internal perceptual noise and not the actual differences related to formulation changes between control and variant samples (O'Mahony and Rousseau 2003). This confusability of samples is relevant to the food industry when product differences used in testing may be very small (d'<0.5 (Ishii and others 2014)).

Within the articulation process of the PA-2-AFC test method, various attributes were identified by subjects as seen in Table 4.5. Only 28.3% of subjects identified the difference between samples as sour. 33.3% of subjects identified the difference between samples as sweet which may viewed as a result of taste mixture suppression and interpreted similarly to sourness, but these results do not explain why performance was lower on the PA-2-AFC method than either Researcher-Designated method.

While subjects were directed to which difference to attune themselves to in the RD-2-AFC method, the PA-2-AFC method lacked this designation and perceptions caused by perceptual noise may have been less easily ignored. The goal of the PA-2-AFC method is to allow a researcher to utilize a specified discrimination testing method when a specific difference is not easily identified. Current results indicate additional training may be needed for panelists to identify and articulate differences between samples which are pertinent to actual formulation differences.

In addition to specifying a difference between samples, the articulation process used in the PA-2-AFC method allows a panelist to declare they cannot detect a difference between pre-test samples after ten sample comparisons. It is interesting to note that while the relative degree of difference between samples used in the study was low ($d'<1$), only one subject stated that they were unable to differentiate between samples. Overall low performance on the PA-2-AFC method indicates that the perceived ability for subjects to differentiate between the samples may be exaggerated and guessing of a perceived difference may actually be the case. These results give evidence to the overconfidence effect (Dunning and others 1990) where subjects rate their decisions with levels of confidence above the actual success rate. Pollack and Decker (1958) observed the overconfidence effect with auditory stimuli in a signal detection experiment. The

researchers found that overconfidence in judgement was higher for more difficult tasks than in easier tasks. Overconfidence in the articulation process of the PA-2-AFC may have led to the poor performance observed for the method.

The assumption made in the PA-2-AFC method is that subjects identify a perceived difference between samples and utilize this difference in completing a standard 2-AFC test with a chance probability of 0.5. While this is true if subjects are actually perceiving a difference between the samples, the assumption fails if subjects resort to guessing during the articulation process. In the PA-2-AFC procedure, subjects first identify the nature of the difference between the samples. They are, then, asked to identify which sample is higher in the specified attribute. By guessing which samples is higher in the specified attribute, a subject has now decreased the chance probability for the test method. The subject must now select a sample which matches their previous guess, essentially doubling the chance of an incorrect response, thus creating a chance level probability of 0.25.

In samples with a large degree of difference, guessing is expected to be low, but in samples with a low degree of difference ($d'<1$) guessing may occur at a higher rate. It is hypothesized that in the case of low degree of difference samples ($d'<1$), the chance level probability for the PA-2-AFC test method may actually be 0.25 as opposed to the designated 0.50 chance probability typically associated with a 2-AFC test.

An increase in power was observed as subjects performed the test methods throughout the study. The test methods utilized in the study were performed in sequence of PA-2-AFC, RD-2-AFC without warm-up, and RD-2-AFC with warm-up. The order in which these tests were performed was selected to prevent subjects from becoming biased by previous exposure to samples and identification of attributes from testing instructions. If the PA-2-AFC method had

been performed after a researcher-designated method, subjects may have been influenced to select the term already provided in the researcher-designated method. Additionally, if the RD-2-AFC without warm-up were performed after the RD-2-AFC with warm-up, prior sample exposure may have induced warm-up effect to the method without warm-up.

While the testing sequence was a necessity for the study design, it does not eliminate the possibility of increased subject performance due to learning effects (McBride and Laing 1979; Frijters and others 1982). As power comparison was the intended goal of the study, utilizing a design with different judges for each method may provide an additional source of variation to the data comparison across different methods (McClure and Lawless 2010; O'Mahony 2013), thus also not free from bias.

4.5 Conclusions

As the 2-AFC method examined in the current study utilized panelists to articulate the nature of the difference, further research may examine if panelist training on identifying sample differences is needed in order to achieve the results predicted by Thurstonian modeling. By including training for subjects similar to that which is often a part of descriptive analysis techniques on identification of attributes which differentiate samples, subjects may increase accuracy in attribute selection and increase performance on PA-2-AFC methods.

It is hypothesized that subjects have a high rate of guessing when selecting which sample is higher in the subject's specified attribute. This guessing would lead to a decrease in the chance probability of a correct response. The range of d' of the sample set used in testing may play a large role in the rate of guessing by subjects. For sample sets with d' values below one, guessing is likely higher than sample sets with higher d' values, as the samples' differences are more likely to be obscured by perceptual noise. Although the degree of difference between samples is

confusable, subjects may still identify an attribute by which to differentiate samples and designate a sample which is higher in the specified attribute for sake of completing the assigned task.

Future research should focus on identifying the rate at which subjects resort to guessing during the specification portion of the articulation process. By doing so, further insight into the effect of d' on the rate of guessing and the chance probability may be garnered. Knowledge of a d' range where guessing is significantly lowered may aid in determining at which level of d' a 0.25 or a 0.50 chance probability is reflective of the true chance probability. By comparing the articulated attributes to known formulation difference between samples, as well as asking subjects their degree of certainty in their selected attribute, a model may be developed to predict the relative rate of guessing at varying d' levels.

4.6 References

Angulo O, Lee H, O'Mahony M. 2007. Sensory difference tests: Overdispersion and warm-up. Food Quality and Preference 18(2):190-5.

Bi J, Ennis DM. 1999. The Power Of Sensory Discrimination Methods Used In Replicated Difference And Preference Tests. J.Sens.Stud. 14(3):289-302.

Byer AJ, Abrams D. 1953. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Dacremont C, Sauvageot F, Duyen TH. 2000. Effect Of Assessors Expertise Level On Efficiency Of Warm-up For Triangle Tests. J.Sens.Stud. 15(2):151-62.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Dunning D, Griffin DW, Milojkovic JD, Ross L. 1990. The overconfidence effect in social prediction. J.Pers.Soc.Psychol. 58(4):568.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis JM, Jesionka V. 2011. The Power Of Sensory Discrimination Methods Revisited. J.Sens.Stud. 26(5):371-82.

Frijters JER, Blauw YH, Vermaat SH. 1982. Incidental training in the Triangular Method. Chemical Senses 7(1):63-9.

Frijters JER, Kooistra A, Vereijken PFG. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Percept.Psychophys. 27(2):176-8.

Gridgeman NT. 1970. A Reexamination of the Two-Stage Triangle Test for the Perception of Sensory Differencesa. J.Food Sci. 35(1):87-91.

Heron WT. 1928. The Warming-Up Effect in Learning Nonsense Syllables. The Pedagogical Seminary and Journal of Genetic Psychology 35(2):219-28.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Quality and Preference 31(0):49-55.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Mata-Garcia M, Angulo O, O'Mahony M. 2007. On Warm-up. J.Sens.Stud. 22(2):187-93.

McBride RL, Laing DG. 1979. Threshold determination by triangle testing: effects of judgemental procedure, positional bias and incidental training. Chemical Senses 4(4):319-26.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M, Thieme U, Goldstein LR. 1988. The Warm-up Effect as a Means of Increasing the Discriminability of Sensory Difference Tests. J.Food Sci. 53(6):1848-50.

O'Mahony M. 2013. The Tetrad Test: Looking Back, Looking Forward. J.Sens.Stud. 28(4):259-63.

O'Mahony M, Rousseau B. 2003. Discrimination testing: a few ideas, old and new. Food Quality and Preference 14(2):157-64.

Pollack I, Decker LR. 1958. Confidence ratings, message reception, and the receiver operating characteristic. J.Acoust.Soc.Am. 30(4):286-92.

Rousseau B. 2003. Sensory Evaluation | Sensory Difference Testing. In: Benjamin Caballero, editor. Encyclopedia of Food Sciences and Nutrition (Second Edition). Oxford: Academic Press. p 5141-7.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

Thieme U, O'Mahony M. 1990. Modifications To Sensory Difference Test Protocols: The Warmed Up Paired Comparison, The Single Standard Duo-trio And The A-not A Test Modified For Response Bias. J.Sens.Stud. 5(3):159-76.

4.7 Tables and Figures

Table 4.1. Composition of samples used in basic taste identification procedures during panel prescreening.

Solution	Concentration (stimuli/water)
Sweet (Sucrose)	20g/980g
Sour (Citric Acid)	0.6g/999.4g
Bitter (Caffeine)	0.7g/999.3g

Table 4.2. Composition of samples used in basic taste intensity ranking procedures during panel prescreening.

Solution	Chemical	g/L solution
Sweet 1	Sucrose	20
Sweet 2		50
Sweet 3		100
Sour 1	Citric acid	0.3
Sour 2		0.6
Sour 3		1.0
Bitter 1	Caffeine	0.4
Bitter 2		0.8
Bitter 3		1.2

Table 4.3. Composition of samples used in aroma identification procedures during panel prescreening.

Solution	Sample Composition
Clear vanilla extract	One mL extract
Orange extract	One mL extract
Lemon extract	One mL extract

Table 4.4. Comparison of proportion of correct responses and d' values for Panelist-Articulated (PA) and Researcher-Designated-2-Alternative Forced Choice (RD-2-AFC) methods with and without warm-up samples. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 indicates the variance of d'

Method	% Correct Responses	d'	σ^2 of d'	Sig diff between d' values
PA-2-AFC	40.0%	0	N/A	N/A
RD-2-AFC without Warm-up	55.7%	0.20	0.05	0.14
RD-2-AFC with Warm-up	68.9%	0.70	0.06	

Figure 4.1. Comparison of power between Panelist-Articulated-2-Alternative Forced Choice (PA-2-AFC) method (n=60) and Researcher-Designated-2-AFC (RD-2-AFC) methods with and without warm-up samples (n=61), alpha=0.05.

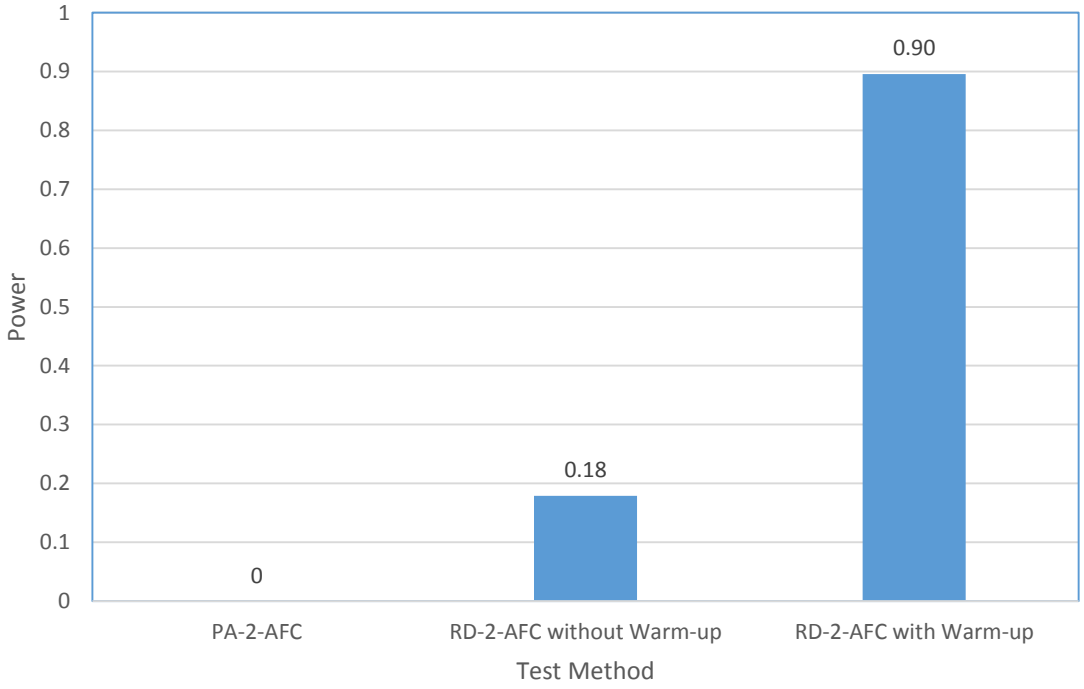


Table 4.5. Terms generated during articulation process of the Panelist-Articulated-2-Alternative Forced Choice method (n=60)

Attribute	Frequency of Articulation
Sweetness	20
Sourness	17
Carbonation	6
Stronger flavor	6
Bitterness	4
Sweet Aftertaste	2
Lightness	1
Lemon flavor	1
Syrupiness	1
Stronger Aftertaste	1
Fruit Flavor	1

Chapter 5: Comparison of Specified and Non-Specified Tetrad and Triad Methods Using Beverage Samples

5.1 Abstract

The tetrad method has garnered popularity in recent years as it has been demonstrated to have higher power than other non-specified test methods. The specified tetrad method has also been stated as having higher theoretical power than other widely-used specified test methods such as the 2-AFC. A comparison of triadic and tetrad methods, both specified and non-specified, has not been conducted using commercial beverages with small perceptual differences between samples. The goal of the study was to compare the power of triadic and tetrad test methods using samples with small sensory differences across a spectrum of beverages available on the market today.

Triangle, 3-AFC, unspecified tetrad, and specified tetrad methods were performed by 60 pre-screened subjects. Tea, vegetable juice, and carbonated beverages were used for comparison of the test method power. In addition to power, the degree of difference between products, estimated using d', was determined and compared to Thurstonian modeling predictions for specified and non-specified test methods.

Low d' samples were found to have higher power with non-specified test methods, contradicting predictions made using Thurstonian modeling. Samples which had higher d', were found to match expected findings based on Thurstonian predictions, with specified test methods resulting in higher power than non-specified methods. Based on these findings, complexity of samples as well as the degree of difference between samples used in testing may impact method power. Future research should be conducted to determine the extent to which sample complexity impacts the results of discrimination testing.

Keywords: Discrimination testing, Triangle, Tetrad, 3-AFC, Specified Tetrad, Power

5.2 Introduction

Having powerful methods is of the upmost importance when conducting sensory discrimination testing. Without sufficient power, sensory testing methods fail to detect important differences consistently and may lead a researcher to miss consumer relevant differences (Bi and Ennis 1999; Ennis and Jesionka 2011). In a time when the food and beverage industry are increasingly focused on the reformulation of existing products, the need for powerful sensory discrimination testing is great.

A typical approach to selecting a powerful test method is to use a method which specifies the nature of the difference between products, such as the 3-AFC method. By utilizing a specified test method, subjects may employ a skimming decision strategy when completing the task as opposed to the comparison of distances strategy used in the triangle method (Frijters 1979; O'Mahony 1992). In the skimming strategy, subjects are believed to select the strongest or weakest sample regardless of the proximity in perceptual distance of other samples. In the comparison of distances strategy, subjects compare the perceptual distance between samples, identifying which samples are closest to one another. The sample which is farthest from the other pair is deemed as odd. The skimming strategy of the 3-AFC method has been shown to increase performance and thus result in a more powerful test (Byer and Abrams 1953; Gridgeman 1970; Frijters 1979; Frijters and others 1980; Ennis 1990; Ennis 1993; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005). However, situations may arise when it is either impossible or unadvisable to specify the nature of the difference between the two samples (Ennis 2013). In these situations, researchers must select among the non-specified test methods.

Of recent development has been the increased attention given to the non-specified test method, the tetrad test (Masuoka and others 1995; Delwiche and O'Mahony 1996; Ennis and Jesionka 2011; Ennis 2012; Ennis 2013; Rousseau and Ennis 2013; O'Mahony 2013; Ennis and Christensen 2014; Ishii and others 2014). The tetrad method has been shown to be a more powerful test method than the triangle test (Ennis 2012; Rousseau and Ennis 2013; Garcia and others 2013; O'Mahony 2013; Ennis 2013; Ennis and Christensen 2014; Ishii and others 2014), which has long been the standard when conducting non-specified discrimination testing. The tetrad method, too, has a specified test counterpart, the specified tetrad test. The specified tetrad method has theoretical and experimental evidence showing that it has a power advantage over the 2-AFC method (Garcia and others 2013), which has been recommended for use as one of the most powerful test methods.

One potential problematic aspect of the tetrad method is the inclusion of a fourth test stimulus. In fatiguing or complex samples, the addition of a fourth test stimulus has the potential to negate the theoretical increased power of the tetrad method due to an increase in perceptual noise (Ennis 2012). In commercial beverages which include high impact flavors, texture components such as fruit or vegetable pulp, as well as carbonation, it remains to be determined if the tetrad method is suitable for testing.

In addition to determining if test methods retain their theoretical advantage in complex stimuli, the relative difference between samples used in testing is also of importance. A measure of determining the relative difference between samples is the d' of the sample set (O'Mahony 1992). The smaller the d' a sample set has, the more confusable the samples are, in terms of sensory perception. Much of the research performed which aims at comparing the power of test methods is performed using samples with a relatively large d' (d'>1) (Masuoka and others 1995;

Rousseau and others 1998; Rousseau and others 1999; Lee and others 2007). When conducting sensory testing to determine if consumers can notice a difference between reformulated samples, samples with a d' near 0.5 may be more relevant (Ishii and others 2014).

The goal of the current study was to compare the proportion of correct responses and power of 3-AFC and specified tetrad methods to triangle and unspecified tetrad test methods using commercial beverages over a range of d'. It was hypothesized that the increased proportion of correct responses for the 3-AFC and specified tetrad methods predicted by Thurstonian modeling will decrease with the use of multidimensional samples such as complex commercial beverages.

5.3 Materials and Methods

5.3.1 Subjects

Prior to acceptance into the study, subjects performed sensory acuity tasks to ensure a base level of sensory acuity for all subjects. The screening tasks were performed in six sections. In the first section, subjects were asked to taste solutions which represented one of the basic tastes and identify the taste sensation. After basic taste identification, subjects were presented with three samples of varying concentration for three different taste sensations, sweet, sour, and bitter. Subjects were instructed to sort the samples in order of intensity from the least intense to the most intense for each of the three taste sensations. In the third section of screening tasks, subjects were presented with three aroma compounds placed on cotton balls in lidded containers and asked to match the perceived aroma to the listed aromas. Next, subjects were instructed to perform two discrimination testing tasks, Triangle and Duo-Trio test methods. Following each discrimination test, subjects were asked how they differentiated samples. After taste, aroma, and

discrimination portions of the screening procedures, subjects identified the frequency of consumption for differing beverage categories and then listed their availability for participation in the testing procedure and demographics. A detailed description of samples utilized in subject screening can be found in Chapter 4 (Bloom 2015).

To be included in the study, subjects needed a minimum of 70% correct responses on taste and aroma samples, be between the ages of 18-55 years, frequent consumers of the product categories matching the categories used during testing and have availability for all scheduled testing sessions. Sixty-one subjects passed screening procedures and were selected as participants in the study out of the 101 subjects screened. After completion of the study, subjects received monetary compensation.

5.3.2 Samples

The products used in testing were divided into three product categories, carbonated beverages, juice, and tea. In each product category, three product comparisons were completed by each subject for each test method. In order to preserve carbonation, carbonated beverages were opened immediately prior to subject evaluation. No carbonated beverage was held for more than two minutes after opening. Samples were served in lidded 60-mL sample cups labeled with random three-digit codes. Samples were served at room temperature (22°C).

Carbonated Beverage Products

Three different commercially available carbonated beverages were selected for use in the study. The first product was a reduced calorie root beer beverage. Control and variant samples differed based on packaging of a commercially available carbonated beverage. Control beverages were packaged in aluminum cans and variant samples were packaged in PET bottles. Formulation differences between the products were the same. The second product was a regular carbonated cola with difference between control and variant samples being in the sweetener used in formulation. The third product was a reduced calorie carbonated cola. Control and variant samples differed based on the sweeteners used in formulation.

Juice Products

The same control sample was used in comparison with three variant samples for a total of three product comparisons of tomato juice. Control and variant samples differed based on saltiness. The control sample was a commercially available, low sodium tomato juice. For the three variations of tomato juice, sodium chloride was added to the control sample to increase the saltiness of the juice. Three levels of sodium chloride were used to create the three variant samples: Juice 1 was the control juice with 41.4 mg of NaCl added per liter, Juice 2 was the control juice with 82.8 mg of NaCl added per liter, Juice 3 was the control juice with 124.3 mg of NaCl added per liter.

Tea Products

A commercially available lemon tea was selected for the creation of control and three variant samples. For the control sample, the tea was used as purchased from a commercial retailer. For the three variant samples, sucrose was added to the control beverage to increase the

sweetness of the product. Three levels of sweetness were created through the use of increasing sucrose additions to create the three variations: Tea 1 was the control tea with 5g of sucrose added per liter, Tea 2 was the control tea with 15g of sucrose added per liter, Tea 3 was the control tea with 25g of sucrose added per liter.

5.3.3 Experimental Procedure

Four test conditions were used: triangle, 3-AFC, unspecified tetrad, and specified tetrad. The tasks completed for each method are listed in Table 5.1. In each testing session, subjects performed one replicate for each of the test methods. The order of test methods was randomized across subjects for each testing session. Between each test method, subjects were given a timed five minute break. All possible sample serving orders were randomized across subjects for each test method. Subjects performed one session for each product comparison for a total of nine testing sessions. Each session lasted approximately 30 minutes.

All tests were conducted in isolated sensory booths, where the temperature was set approximately at 22°C and relative humidity at 33%. Tests were performed under incandescent lighting as no visual differences were observed between samples. Subjects were instructed to rinse before testing and between samples with carbonated water, warm water, and room temperature water to reduce interstimulus carry-over effects. Data were collected using Compusense® *five Plus* (Version 5.6: Guelph ON, Canada).

5.3.4 Data Analysis

Chi-Square analysis was performed to determine if significant difference existed between d' values produced from different test methods of the same product comparison (Table 5.2-5.4).

Test methods which resulted in a d' of 0 were not included in Chi-Square analysis. Within each product comparison, methods which resulted in a d' of 0 were compared to the highest d' from one of the other test methods. To determine if the highest d' differed significantly from 0, z-values were calculated and compared at an alpha level of 0.05 (Bi and others 1997). Results can be found in Table 5.5. Additionally, power was determined for each test method and product comparison using d' estimated from the number of correct responses for each test, a sample size of 61, and an alpha level of 0.05. Data analysis was performed using IFPrograms™ version 8.11 (The Institute for Perception, USA).

5.4 Results and Discussion

The proportion of correct responses, d', and power values for each method can be found in Tables 5.6-5.8. When comparing methods, a focus was placed on difference between power and sensory difference between product pairs used in testing which we estimated by calculating d'. As the specified tetrad method has a 1/6 guessing probability compared to the 1/3 guessing probability of the triangle, 3-AFC, and unspecified tetrad methods a comparison of proportion of correct responses between the specified tetrad method and other methods is not appropriate (Garcia and others 2013).

The methods which resulted in the highest power differed by product type and the level of d' of samples used in testing. For the carbonated beverage and juice products, the, low d' values were observed and reflect the types of samples that may be used in consumer discrimination testing (Ishii and others 2014). The tea products used in testing had much higher d' values, which resemble the d' of samples commonly used in studies within literature

mentioned previously (Masuoka and others 1995; Rousseau and others 1998; Rousseau and others 1999; Lee and others 2007).

Carbonated Beverage Products

Results for the carbonated beverage samples used in testing can be found in Table 5.6. Overall, low d' (d'<0.5) was observed for diet soda beverage products used in testing. Based on the number of subjects used in testing, low power was observed for all test methods. For the diet soda product, only the triangle test method was able to produce results with a d' above 0 (d'=0.43). When comparing this d' value to 0 (Table 5.5), no significant difference was determined.

For the diet-root-beer carbonated beverage, the triangle test method resulted in the highest d' (d'=0.43) (Table 5.6) . Only the unspecified tetrad method resulted in a d' of 0. No significant differences were observed between the d' values from any of the test methods when diet-root-beer beverages were tested.

For the regular soda carbonated beverage, the unspecified tetrad method resulted in the largest d' value between samples. Both the 3-AFC and specified tetrad methods resulted in d' of 0. There was a significant difference between the d' of 0.84 produced by the unspecified tetrad method and 0 by other methods (Table 5.5). When the proportion of correct responses was compared between the 3-AFC, triangle, and unspecified tetrad method, there was no increase in correct responses for the 3-AFC method as predicted by Thurstonian modeling (Frijters 1979). Both the triangle and unspecified tetrad methods resulted in a higher proportion of correct responses than the 3-AFC method.

Although low power was observed for the carbonated beverage category, the results provide interesting findings which lead to a breakdown of the assumptions present in Thurstonian modeling predictions. Significant differences were observed among d' values generated by different methods for the regular carbonated beverage product. Ennis (1998) has suggested that significant differences observed between d' values of test methods may indicate sample multidimensionality or sample order effects. As test order and sample presentations were randomized across subjects, the differences in d' values may signify dimensionality impacting the test results in this study.

Juice Products

The proportion of correct responses, d', and power for each product comparison can be found in Table 5.7. For the Juice 1 product comparison, the specified tetrad method was the only method to produce a d' value above 0. When compared to a d' of 0 (Table 5.5), the d' of 0.42 generated by the specified tetrad method significantly differ from the other test methods. The specified tetrad method also resulted in the highest power for this product comparison (power=0.658).

For the Juice 2 product comparison, the 3-AFC method was the only test method which resulted in a d' of 0. The triangle test method resulted in the highest d', 0.97, followed by the unspecified tetrad method, 0.51. Between the 3-AFC, triangle, and unspecified tetrad method, the 3-AFC also resulted in the lowest proportion of correct responses. These results do not follow expected results predicted by Thurstonian modeling (Frijters 1979). When comparing the d' results for the 3-AFC (d'=0) and triangle test (d'=0.97), significant difference exist between the methods (alpha=0.05) (Table 5.5).

When comparing the results from test methods for the Juice 3 product comparison, the triangle test method resulted in the largest d' value (d'=0.36) but this d' does not differ significantly from 0. Both specified and unspecified tetrad methods resulted in a d' of 0, which may indicate that the addition of the fourth test stimulus may lower the sensitivity of the test methods (Ennis 2012), when using a food product that may result in more sensory fatigue like the tomato juice product.

The comparison of methods using low d' samples is an important addition to the literature available. As Ishii and others (2014) discussed, samples with a d' around 0.5 are at a level of difference relevant to consumer testing. A triangle test conducted using 60 subjects nears a p-value of 0.05 at a d' of approximately 1.0. Conducting discrimination tests using samples with a d' above 1.0 or in some instances near 2.0 or 3.0 are not relevant to a majority of discrimination testing scenarios with commercial products. One would not utilize resources to confirm differences that are already obvious to be detected by consumers.

The Juice products used differed in the amount of sodium chloride between control and variant samples for each product comparison. Juice 1 had the lowest addition of sodium chloride between control and variant, with Juice 2 and Juice 3 having increasing additions of sodium chloride. The expected outcome of these additions was an increase in d' for methods. Juice 1 presented with the control juice was expected to have the lowest d' values and Juice 3 compared with the control was expected to have the largest d'. Experimentally determined d' values do not follow this trend (Table 5.7), as Juice 2 resulted in the highest d' values between samples across test methods.

Differences between expected d' trends and observed values may indicate that the addition of sodium chloride affected the sensory perception of the samples in ways other than

merely a difference in saltiness perception. Gillette (1985), when looking at several food systems including tomato soup, found that the presence of NaCl in foods not only enhanced saltiness perception but also mouthfeel, sweetness, and balance as well as decreased off-notes. Even the change of one ingredient may have multidimensional impacts to sensory perception. These multidimensional changes to sensory perception may cause the deviation from expected results observed in the juice products tested.

Tea Products

Proportion of correct responses, d', and power results for the tea products used in testing can be found in Table 5.8. When comparing test methods for the Tea 1 product, no significant differences were observed among the d' values generated by the different tests (Table 5.4). Power for the specified methods (3-AFC = 1.000, specified tetrad = 0.990) was much greater than the non-specified methods (triangle = 0.179, unspecified tetrad = 0.114), again confirming Thurstonian predictions (Frijters 1979).

When Tea 1 was presented with the control beverage, a reduction in d' and power was found between triadic and tetradic methods (Table 5.8). One possible explanation for the reduction in power is that the introduction of a fourth stimulus used in tetradic methods has the possibility of reducing test sensitivity (Ennis 2012). As the level of sucrose added to the variant Tea products increased the effects of the fourth stimulus diminished. When Tea 2 and Tea 3 were presented with the control beverage there was no reduction in d' or power. These results suggest that the reduction in sensitivity caused by the addition of a fourth test stimulus in tetradic methods may be d' dependent. Previous research has focused on the fatiguing nature of samples as a reason for selecting triangular methods over tetradic methods (Ennis 2012; Carlisle 2014;

Ennis and Christensen 2014), but the relative degree of difference between samples may also be important.

Tea products differed in the amount of sucrose added between control and variant formulations with Tea 1 having the lowest level of sucrose added and Tea 3 having the highest level of sucrose added. It was expected that Tea 1 would have the lowest d' values and Tea 3 would have the largest. Based on the observed experimental results, the trend in d' values follow these expectations. No significant differences were observed between d' values obtained from the test methods for any of the tea products.

As predicted by Thurstonian modeling (Frijters 1979; Ennis and Jesionka 2011), the 3-AFC method had a higher proportion of correct responses compared to the triangle and unspecified tetrad methods. It is important to note the level of significance resulting from methods in the Tea product category. When presented with the control beverage, Tea 1, Tea 2, and Tea 3 had relatively high d' values (Table 5.8). For the triangle method, the d' between control and Tea 3 was 2.23. When presented with the control beverage, Tea 2 and Tea 3 products resulted in highly significant findings (p-value<0.001). One can imagine that samples with differences this significant can hardly be considered confusable. Conducting sensory discrimination testing using sample pairs of this difference would most likely be to confirm difference as opposed to identifying difference. Although it may not be possible to determine the d' of a sample set prior to testing, these samples are distinct in their difference from one another and are not suitable for discrimination testing.

5.5 Conclusions

If method comparison studies are conducted routinely with large d' sample sets, relevance to actual consumer testing scenarios may be missed. As observed in the findings from this study, the comparison of test methods using low d' samples becomes more difficult as deviations from expected theoretical results occur frequently. While not easy to explain all findings, the present study provide instances for which method selection is not cut and dry.

Comparison of discrimination testing methods with in-house panels may be important before converting the test methods used to conduct day to day testing. As demonstrated by the findings presented in the current study, the complexity of samples and how they differ may cause test method power to deviate from those found in literature. By using the product categories important to each individual business and with the degree of difference between samples common to testing, sensory scientists may identify methods of high power for use in business applications. Additionally, through in-house testing a common degree of difference can be found through the estimation of d' which may be useful in determining the proper sample size needed for discrimination testing with high power. While literature does provide a basis for selection based on years of work exploring the use of Thurstonian modeling, with the variability found in human subject performance and the wide range of products used in discrimination testing with the food and beverage industries there may not be a method that is "one-method-fits-all-test-scenarios".

It is hypothesized that formulation dimensionality as well as the level of d' impact whether specified or non-specified methods results in greater power. As current findings are limited to the commercial makeup of samples used in testing, future research should explore the

influence of both d’ and dimensionality to determine their effects on method power in a controlled beverage system.

5.6 References

Bi J, Ennis DM. 1999. The Power Of Sensory Discrimination Methods Used In Replicated Difference And Preference Tests. J.Sens.Stud. 14(3):289-302.

Bi J, Ennis DM, O'Mahony M. 1997. How to estimate and use the variance of d' from difference tests. J.Sens.Stud. 12(2):87-104.

Bloom DJ. 2015. Sensory Discrimination Testing Methodology Selection Based on Beverage Complexity.

Byer AJ, Abrams D. 1953. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Carlisle SL. 2014. Comparison of Triangle and Tetrad Discrimination Methodology in Applied, Industrial Manner. [dissertation]. Knoxville: University of Tennessee.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM. 1998. Thurstonian Scaling for Difference Tests. IFPress 1(3):2.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Ennis JM. 2013. The Year of the Tetrad Test. J.Sens.Stud. 28(4):257-8.

Ennis JM. 2012. Guiding The Switch From Triangle Testing To Tetrad Testing. J.Sens.Stud. 27(4):223-31.

Ennis JM, Christensen RHB. 2014. Precision of measurement in Tetrad testing. Food Quality and Preference 32, Part A(0):98-106.

Ennis JM, Jesionka V. 2011. The Power Of Sensory Discrimination Methods Revisited. J.Sens.Stud. 26(5):371-82.

Frijters JER. 1979. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER, Kooistra A, Vereijken PFG. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Percept.Psychophys. 27(2):176-8.

Garcia K, Ennis JM, Prinyawiwatkul W. 2013. Reconsidering the Specified Tetrad Test. J.Sens.Stud. 28(6):445-9.

Gillette M. 1985. Flavor effects of sodium chloride. Food technology (USA)

Gridgeman NT. 1970. A Reexamination of the Two-Stage Triangle Test for the Perception of Sensory Differencesa. J.Food Sci. 35(1):87-91.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Quality and Preference 31(0):49-55.

Lee H-, van Hout D, Hautus MJ. 2007. Comparison of performance in the A–Not A, 2-AFC, and same–different tests for the flavor discrimination of margarines: The effect of cognitive decision strategies. Food Quality and Preference 18(6):920-8.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Masuoka S, Hatjopoulos D, O'Mahony M. 1995. Beer Bitterness Detection: Testing Thurstonian And Sequential Sensitivity Analysis Models For Triad And Tetrad Methods. J.Sens.Stud. 10(3):295-306.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M. 1992. Understanding Discrimination Tests: A User-friendly Treatment Of Response Bias, Rating And Ranking R-index Tests And Their Relationship To Signal Detection. J.Sens.Stud. 7(1):1-47.

O'Mahony M. 2013. The Tetrad Test: Looking Back, Looking Forward. J.Sens.Stud. 28(4):259-63.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivit of the same-different test: Comparison with triangle and duo-trio methods. J.Sens.Stud. 13(2):149-73.

Rousseau B, Ennis JM. 2013. Importance of Correct Instructions in the Tetrad Test. J.Sens.Stud. 28(4):264-9.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-84.

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

5.7 Tables and Figures

Table 5.1. Sensory discrimination testing methods utilized in testing for Aim 3

Test Method	Task	Sample Presentation Orders	Chance Probability
Triangle	Identify the odd sample	AAB, ABA, BAA, BBA, BAB, ABB	1/3
Researcher-Designated-3-AFC	Identify the stronger sample	AAB, ABA, BAA, BBA, BAB, ABB	1/3
Unspecified Tetrad	Group the samples into 2 groups of 2 based on similarity	AABB, ABAB, ABBA, BBAA, BABA, BAAB	1/3
Specified Tetrad	Identify the two stronger samples	AABB, ABAB, ABBA, BBAA, BABA, BAAB	1/6

Table 5.2. Comparison of d' values from different test methods for using carbonated beverage samples. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 is the variance associated with d' for each test method. Test methods with a d' equal to 0 were not included in Chi-Square analysis. Bolded p-values are below 5% indicating that d' values between methods are determined to be significantly different.

Product	Test	d'	σ^2 of d'	chi-square	critical	p-value
Diet Soda	3-AFC	0.00	N/A	N/A	N/A	N/A
	Triangle	0.43	0.648			
	Unspecified Tetrad	0.00	N/A			
	Specified Tetrad	0.00	N/A			
Regular Soda	3-AFC	0.00	N/A	0.236	3.841	0.627
	Triangle	0.43	0.648			
	Unspecified Tetrad	0.84	0.064			
	Specified Tetrad	0.00	N/A			
Diet-root-beer	3-AFC	0.23	0.045	0.181	5.991	0.913
	Triangle	0.43	0.648			
	Unspecified Tetrad	0.00	N/A			
	Specified Tetrad	0.14	0.044			

Table 5.3. Comparison of d' values from different test methods for using juice samples. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 is the variance associated with d' for each test method. Test methods with a d' equal to 0 were not included in Chi-Square analysis. Bolded p-values are below 5% indicating that d' values between methods are determined to be significantly different.

Product	Test	d'	σ^2 of d'	chi-square	critical	p-value
Juice 1	3-AFC	0.00	N/A	N/A	N/A	N/A
	Triangle	0.00	N/A			
	Unspecified Tetrad	0.00	N/A			
	Specified Tetrad	0.42	0.04			
Juice 2	3-AFC	0.00	N/A	3.058	5.991	0.217
	Triangle	0.97	0.179			
	Unspecified Tetrad	0.51	0.134			
	Specified Tetrad	0.17	0.045			
Juice 3	3-AFC	0.04	0.048	0.102	3.841	0.749
	Triangle	0.36	0.952			
	Unspecified Tetrad	0.00	N/A			
	Specified Tetrad	0.00	N/A			

Table 5.4. Comparison of d' values from different test methods for using tea samples. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 is the variance associated with d' for each test method. Test methods with a d' equal to 0 were not included in Chi-Square analysis. Bolded p-values are below 5% indicating that d' values between methods are determined to be significantly different.

Product	Test	d'	σ^2 of d'	chi-square	critical	p-value
Tea 1	3-AFC	1.24	0.049	3.132	7.815	0.372
	Triangle	0.76	0.247			
	Unspecified Tetrad	0.43	0.336			
	Specified Tetrad	0.82	0.036			
Tea 2	3-AFC	1.43	0.053	8.415	7.815	0.089
	Triangle	1.64	0.108			
	Unspecified Tetrad	1.19	0.047			
	Specified Tetrad	1.9	0.036			
Tea 3	3-AFC	1.91	0.067	3.267	7.815	0.352
	Triangle	2.23	0.105			
	Unspecified Tetrad	1.59	0.044			
	Specified Tetrad	1.97	0.045			

Table 5.5. Methods with a d' = 0 were compared to the highest d' value from a method using the same products to determine if significant differences existed between d' values from difference methods for the same products. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. σ^2 is the variance associated with d' for each test method. Bolded p-values are below 5% indicating that the d' differs significantly from 0.

Product	Test	d'	σ^2 of d'	z score	p-value
Diet Soda	Triangle	0.43	0.648	0.534	0.295
Regular Soda	Unspecified Tetrad	0.84	0.064	3.309	**0.001**
Diet-root-beer	Triangle	0.43	0.648	0.534	0.295
Juice 1	Specified Tetrad	0.42	0.040	2.093	**0.018**
Juice 2	Triangle	0.97	0.179	2.295	**0.011**
Juice 3	Triangle	0.36	0.952	0.369	0.359

Table 5.6. Summary of discrimination testing results from methods using carbonated beverage samples. P value listed designates the significance at which the test method can declare the sample pair different. d’ is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. Power is the probability of declaring a difference when one exists between samples.

Product	Test	Chance Probability	# Correct	Total #	P-value	d'	Power
Diet Soda	3-AFC	0.33	20	60	0.548	0.00	0.040
	Triangle	0.33	21	60	0.440	0.43	0.070
	Unspecified Tetrad	0.33	16	60	0.893	0.00	0.040
	Specified Tetrad	0.17	8	60	0.804	0.00	0.034
Regular Soda	3-AFC	0.33	20	60	0.548	0.00	0.040
	Triangle	0.33	21	60	0.440	0.43	0.070
	Unspecified Tetrad	0.33	27	60	0.040	0.84	0.557
	Specified Tetrad	0.17	10	60	0.554	0.00	0.034
Diet-root-beer	3-AFC	0.33	24	60	0.169	0.23	0.255
	Triangle	0.33	21	60	0.440	0.43	0.070
	Unspecified Tetrad	0.33	17	60	0.831	0.00	0.040
	Specified Tetrad	0.17	12	60	0.292	0.14	0.128

Table 5.7. Summary of discrimination testing results from methods using juice samples. P value listed designates the significance at which the test method can declare the sample pair different. d’ is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. Power is the probability of declaring a difference when one exists between samples.

Product	Test	Chance Probability	# Correct	Total #	P-value	d'	Power
Juice 1	3-AFC	0.33	19	58	0.586	0.00	0.045
	Triangle	0.33	16	58	0.858	0.00	0.045
	Unspecified Tetrad	0.33	14	58	0.951	0.00	0.045
	Specified Tetrad	0.17	16	58	0.025	0.42	0.658
Juice 2	3-AFC	0.33	19	58	0.586	0.00	0.045
	Triangle	0.33	24	58	0.124	0.97	0.340
	Unspecified Tetrad	0.33	22	58	0.270	0.51	0.045
	Specified Tetrad	0.17	12	58	0.252	0.17	0.206
Juice 3	3-AFC	0.33	20	58	0.475	0.04	0.066
	Triangle	0.33	20	58	0.475	0.36	0.067
	Unspecified Tetrad	0.33	19	58	0.586	0.00	0.045
	Specified Tetrad	0.17	9	58	0.647	0.00	0.050

Table 5.8. Summary of discrimination testing results from methods using tea samples. P value listed designates the significance at which the test method can declare the sample pair different. d’ is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by binomial calculations. Power is the probability of declaring a difference when one exists between samples.

Product	Test	Chance Probability	# Correct	Total #	P-value	d'	Power
Tea 1	3-AFC	0.33	42	60	0.000	1.24	1.000
	Triangle	0.33	23	60	0.244	0.76	0.179
	Unspecified Tetrad	0.33	22	60	0.336	0.43	0.114
	Specified Tetrad	0.17	24	60	0.000	0.82	0.990
Tea 2	3-AFC	0.33	45	60	0.000	1.43	1.000
	Triangle	0.33	32	60	0.001	1.64	0.923
	Unspecified Tetrad	0.33	33	60	0.000	1.19	0.952
	Specified Tetrad	0.17	45	60	0.000	1.90	1.000
Tea 3	3-AFC	0.33	51	60	0.000	1.91	1.000
	Triangle	0.33	39	60	0.000	2.23	0.999
	Unspecified Tetrad	0.33	40	60	0.000	1.59	1.000
	Specified Tetrad	0.17	46	60	0.000	1.97	1.000

Chapter 6: Sample Dimensionality Effects on d' and Proportion of Correct Responses in Discrimination Testing

6.1 Abstract

Products in the food and beverage industry have varying levels of dimensionality ranging from pure water to multicomponent food products which can modify sensory perception and possibly influence discrimination testing results. The objectives of the study were to determine the impact of 1) sample dimensionality and 2) complex formulation changes on the d' and proportion of correct response from 3-AFC and Triangle methods.

Two experiments were conducted using 47 prescreened subjects who performed either triangle or 3-AFC test procedures. In Experiment I, subjects performed 3-AFC and triangle tests using model solutions with different levels of dimensionality. Samples increased in dimensionality from one dimensional sucrose in water to three dimensional sucrose, citric acid, and flavor in water. In Experiment II, subjects performed 3-AFC and triangle tests using three dimensional solutions. Sample pairs differed in all three dimensions simultaneously to represent complex formulation changes. Two forms of complexity were compared: dilution, where all dimensions decreased in the same ratio, and compensation, where a dimension was increased to compensate for a reduction in another.

A reduction of the proportion of correct responses was seen for both test methods between one and two dimensional samples. No reduction in correct responses was observed between two and three dimensional samples, which may be a result of odor-taste interaction. No significant differences were demonstrated between methods when samples with complex formulation changes were tested. Results reveal an impact on proportion of correct responses

due to sample dimensionality and should be explored further using a wide range of sample formulations.

Keywords: Discrimination testing, Triangle, 3-AFC, Dimensionality, Complexity, Power, d'

6.2 Introduction

Research conducted in the area of methodology comparison is often performed using model solutions such as sodium chloride in water (Byer and Abrams 1953; O'Mahony and Odbert 1985; Tedja and others 1994; Mata-Garcia and others 2007), or with foods in which one ingredient alteration has been made during formulation (Stillman 1993; Masuoka and others 1995; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Rousseau and others 1998). Within the literature, complex sample alterations are typically studied using samples in which a dilution has been made between control and variant samples (McClure and Lawless 2010; Ishii and others 2014). While a sample dilution may change multiple aspects of a product, it may not match the multidimensionality change caused when a sample undergoes reformulation in the industry, since the removal of an ingredient in formulation is typically compensated for by utilizing some sort of replacement. For example, reduction of fat in salad dressing is typically compensated for by using a texturizing agent. These types of changes to food products may cause unknown sensory perceptual changes leading to increased sample dimensionality.

When multidimensional differences exist between samples assumptions made in describing sample perception in discrimination testing may break down (Ennis and others 2013). These assumptions include the stability of d', or the discriminable distance between sample perceptual means (O'Mahony and others 1994). For this reason, multidimensional models for

discrimination testing have been developed (Ennis and Mullen 1986b). Multidimensional models predict that with increased dimensionality the proportion of correct responses also decreases when dimensions are independent (Ennis and Mullen 1986c). The decrease in proportion of correct responses with increasing dimensionality is believed to be caused by the increase in variance between stimuli (Ennis and Mullen 1985). When correlation between dimensions exist, the proportion of correct responses for a particular method is not only dependent on dimensionality, but also on the relative correlation of dimensions. While model predictions exist for multidimensional samples (Ennis and Mullen 1986a; Ennis and Mullen 1986c), there is little experimental research available exploring the impact of multidimensional samples on the power of the discrimination tests.

The interaction of taste and odor modalities has been explored for intensity ratings of model solutions. Suppression or enhancement induced by taste and odor mixtures is not caused by chemical interactions between the compounds but rather the converging of sensory information in multimodal neurons (Small and Prescott 2005). This convergence of information can lead to an integration and increased sensory perception of the compounds utilized in solution (Dalton and others 2000). Frank and others (1993) found an increase in the sweetness perception of sucrose and odor combinations, but this effect was dependent on the task given to subjects. When subjects were asked to rate only the intensity of sweetness, sweetness ratings increased in the presence of strawberry odor. When subjects were asked to rate all perceptions on different scales, the sweetness ratings were not enhanced.

Findings from intensity rating studies are relevant to discrimination testing comparisons as discrimination tests have similar differences in the tasks for which subjects are asked to perform. Generally, sensory discrimination testing methods are categorized by the task subjects

are asked to perform (Lawless and Heymann 1999). The task which subjects are asked to perform impacts how the test is approached and which mind strategy is used when completing the task (O'Mahony and Rousseau 2003).

In specified testing methods, such as the 2-AFC and 3-AFC methods, subjects are asked to identify the strongest sample on a specific attribute. In these test methods subjects are believed to take on a skimming mind strategy and select the most intense sample for the specified attribute (O'Mahony and others 1994). Non-specific discrimination tests, such as the triangle test method, do not specify an attribute and subjects are thought to take on a comparison of distances strategy by selecting the sample which is the most different in perceptual distance from the other samples present (O'Mahony and others 1994). The differences in the mind strategy employed in a discrimination testing has shown to cause differences in subject performance with specified test methods resulting in a greater proportion of correct responses than non-specified test methods (Frijters 1979; Frijters and others 1980; Ennis 1990; Stillman 1993; O'Mahony and others 1994; Tedja and others 1994; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Liggett and Delwiche 2005).

As both mind strategies used in discrimination testing and the dimensionality of samples have shown to have an impact on sensory perception and performance in testing, the goal of this study was to assess the impact of sample dimensionality and complex formulation changes on sensory discrimination testing performance. It was hypothesized that a reduction of correct responses would occur with an increase in dimensionality for both specified and non-specified test methods. In Experiment I, the experimental proportion of correct responses at three levels of dimensionality was compared to theoretical predictions made by Ennis and Mullen (1985; 1986a) at the calculated d' of the sample set. Additionally, in Experiment II, a comparison of d'

and power for samples with complex formulation investigated in two forms, dilution and compensation. Dilution samples had a reduction in all ingredients in equal ratio between control and variant samples. Compensation samples differed in the same ratio of change as dilution samples for each dimension, but while two dimensions decreased between control and variant, one dimension increased to compensate for the reduction in other dimensions.

6.3 Materials and Methods

6.3.1 Experiment I

6.3.1.1 Subjects

Subjects performed screening procedures prior to acceptance into the study. Screening procedures consisted of taste and aroma identification, taste intensity ranking, performance on triangle and duo-trio discrimination testing procedures, and listing of demographic information and availability. A detailed explanation of screening procedures can be found in Chapter 4 (Bloom 2015). Screening procedures were intended to ensure a minimum level of sensory acuity for comparisons of data from each test group. A minimum of 70% correct responses was required to participate in the study. Out of 70 subjects who attempted screening procedures, 48 subjects met minimum acuity requirements. All subjects were between the ages of 18 and 55 years of age and had no food related allergies. Of the subjects who participated in the study, 32 were female and 15 were male. The 48 subjects were randomly distributed into two test groups of 24 each. One subject in the 3-AFC test group dropped from the study, therefore, results reflect a total of 23 subjects for the test group. Upon completion of the study, subjects received monetary compensation. The same subjects participated in both Experiments 1 and 2.

6.3.1.2 Samples

Model beverage solutions were created based on a formulation found in Maurice Shachman's, "The Soft Drinks Companion" (Shachman 2005). Modifications to the formulation were made to fit the needs of the current study. Three control and variant pairs were created, each pair representing a level of sample dimensionality. One dimensional samples consisted of sucrose in water. Two dimensional samples consisted of sucrose, citric acid in water. Three dimensional samples consisted of sucrose, citric acid, and lemon-lime flavor in water. Each sample pair formulation differed in the amount of sucrose present between control and variant. Full sample formulations can be found in Table 6.1. Reverse osmosis filtered water was used in sample creation. All samples were served in lidded, 60 mL clear plastic cups (Dart Container Corporation, Mason, MI) labeled with random three-digit codes.

Preliminary testing was conducted to ensure confusability of products used in actual testing. A total of 27 subjects were recruited for some parts of the preliminary testing. Of the 27 subjects who participated in preliminary testing, 18 were female and 9 were male. Several subjects participated in multiple preliminary test sessions. Fifteen subjects participated in each product comparison and performed four replicate triangle tests to estimate the d' of the sample set. A total of 14 preliminary testing sessions were conducted over the course of several weeks. Sample sets with a d' near 1.5 were moved forward for inclusion in the actual testing procedures.

6.3.1.3 Experimental Procedure

To avoid the possibility of subjects changing decision strategies during the experiment and to reduce fatigue, one test group was designated to each testing method and performed either

triangle or 3-AFC test methods throughout the experiment. A between groups comparison has been successfully employed within literature to avoid fatigue (Rousseau and others 1999) and changes in decision strategies (McClure and Lawless 2010). Subjects were screened to meet a base level of sensory acuity and decrease variability between judges in each test group. While comparisons between methods have been made, the primary objective was to assess the impact of dimensionality on each method and variability between judges was less of a concern than changes in decision strategies.

For each test session, subjects performed four replicate tests with a two-minute break between each test. One control and variant pair were utilized per testing session for a total of three test sessions per subject. Dimensionality of samples was randomized across subjects to limit possible bias caused by test order effects. As control and variant samples differed based on the amount of sucrose utilized in formulation, for 3-AFC procedures subject were asked to identify the sweeter sample.

Testing was conducted in sensory booths with environmental conditions of 22°C and relative humidity at 33%. Subjects performed testing under incandescent lighting as no visible differences were apparent between samples. Subjects were instructed to rinse prior to beginning testing and between each test sample. The rinse protocol used throughout testing was warm water (43-49°C) followed by room temperature water (22°C). Data were collected using Compusense® *five Plus* (Version 5.6: Guelph ON, Canada).

6.3.1.4 Data Analysis

Binomial analysis was conducted between each replicate test. As replicated test were performed, beta-binomial analysis (Ennis and Bi 1998) was conducted using IFPrograms™

version 8.12 (The Institute for Perception, USA) to estimate overdispersion. For comparison purposes, d' was determined using tables derived by Ennis (1993) and its variance was calculated using tables from Bi and others (1997) as described by Bi and Ennis (1998). Power for each test and product combination was calculated using the calculated d', an alpha of 0.05, the respective number of trials and replications and calculated gamma values using the IFPrograms™ software. Chi-Square analysis was performed to determine if significant differences existed between d' values produced from different sample dimensionalities.

6.3.2 Experiment II

6.3.2.1 Subjects

Subjects utilized in Experiment II were the same subjects which participated in Experiment I. As the3-AFC testing procedures identify sample differences, subjects in the 3-AFC group had knowledge of possible future product formulation changes which may impact decision strategies used in triangle procedures. To minimize these effects, subjects performed the same testing procedures as those in Experiment I.

6.3.2.2 Samples

Sample formulations for Experiment II can be found in Table 6.2. Dilution samples were created with an equal reduction of all non-water sample ingredients between control and variant samples. Samples were created with an 8:7 dilution ratio between control and variant. Compensation samples were created with the same 8:7 ratio change between control and variant samples, but instead of all formulation changes being reduced between control and variant, ratios

of citric acid were increased to compensate for the reduction in flavor used in formulation. Sample preparation and presentation was the same as that utilized in Experiment I.

Preliminary testing was again conducted to ensure that the samples used in testing fell within a confusable range based on d'. Preliminary testing procedures were the same as those listed in Experiment I. Over nine preliminary testing sessions, 22 subjects participated, 15 of which were female and 7 male.

6.3.2.3 Experimental Procedure

Two testing sessions were completed by each subject. Subjects performed four replicated discrimination test with a two-minute break between each replicate. Sample presentation was randomized across subjects and replications. Half of the subjects began with compensation samples and half of subjects began with dilution samples in order to randomize the dilution and compensation treatment effects. Testing was conducted in isolated sensory booths. Subjects were asked to rinse their mouths with warm water (43-49°C) followed by room temperature water (22°C) prior to the first sample and between samples. For 3-AFC testing, subjects were asked to identify the sweeter sample. For triangle testing, subjects were asked to identify the odd sample. Compusense® *five Plus* (Version 5.6: Guelph ON, Canada) was used to collect data.

6.3.2.4 Data Analysis

As in Experiment I, binomial analysis was conducted between each replicate test and beta-binomial analysis was performed on combined replicate data using IFPrograms™ version 8.11 (The Institute for Perception, USA). Power was also calculated using IFPrograms™ software as described in Experiment I.

6.4 Results and Discussion

6.4.1 Experiment I

The probability value based on the beta-binomial model, d' and its variance, power, and overdispersion values can be found in Table 6.3. As hypothesized, the d' of samples for both the triangle and 3-AFC methods decreased with an increase in dimensionality. For the triangle test method, the increase in dimensionality from one dimensional sample (d' = 2.16) to two dimensional sample (d' = 1.29) resulted in a significant decrease in d' ($p = 0.04$). While the same trend followed for the 3-AFC test method, the results were not significant (Figure 6.2).

The observed reduction in the proportion of correct responses with increasing sample dimensionality supports the theoretical predictions made by Ennis and Mullen (1985). When the proportion of correct responses for the triangle test method are compared to simulation predictions made by Ennis and Mullen (1986b) found in Table 6.4, the experimental proportion of correct responses is nearly equivalent to theoretical predictions for one dimensional samples, 63.5% and 64.1%, respectively. If we assume that the added dimensions used in testing do not impact d' as usage rates for both citric acid and flavor are the equivalent in control and variant samples, we can compare the proportion of correct responses for samples with added dimensions to theoretical predictions. For two dimensional samples, experimental proportion of correct responses is nearly 14% lower than theoretical predictions. Smaller differences exist between theoretical and experimental proportion of correct responses for three dimensional samples with the experimental proportion of correct responses being 6.8% lower than theoretical predictions.

While a decrease in d' (Table 6.3) and proportion of correct responses (Figure 6.1) was observed between one dimensional and two dimensional samples, the addition of flavor to create

three dimensional samples did not decrease the d' of samples used in either triangle or 3-AFC testing. The differences observed between theoretical predictions and experimental results found in this study may indicate that the makeup of the added dimension, while not impacting the overall difference from a formulation standpoint, is providing both additional variance and added information to sensory perception (Ennis and Mullen 1985).

As noted previously, odor-taste interactions have been shown to increase perceived sweetness (Frank and others 1993). In 3-AFC procedures, subjects were asked to identify the sweeter sample. It is possible that the odor-taste interaction between the lemon-lime flavor and sucrose found in solution caused an increase in the proportion of correct response. As neither Triangle nor 3-AFC methods resulted in a decrease in correct responses from two to three dimensional samples, it appears the odor-taste interaction occurred for both methods. This supports the findings of Frank and others (1993), who observed interaction between strawberry odor and sweetness, but did not find significant interaction with lemon odor-sweetness mixtures and instructions. It is possible that interactions between mind strategy and product dimensionality may be formulation specific.

Throughout testing, the 3-AFC method resulted in a higher proportion of correct responses than the triangle method (Figure 6.1). This finding supports Thurstonian modeling predictions (Frijters 1979). No significant differences (alpha=0.05) existed between the d' values from different methods using the same products. As no significant differences were observed in the d' values of samples when subjects performed the 3-AFC method, results may indicate a reduction in perceptual noise when using the specified 3-AFC method. Identification of an attribute to differentiate samples may attune subjects to the dimension which has the largest perceptual difference and reduce the noise created by the added sample dimensions. The triangle

test method does not specify an attribute by which subjects are to differentiate samples, which may allow subjects to attune themselves to perceptual noise of dimensions with no actual formulation differences.

6.4.2 Experiment II

Summary data for Experiment II can be found in Table 6.5. No significant differences (alpha = 0.05) existed between the d' values of dilution and compensation samples for either method. Additionally, no significant differences was identified between the d' values of triangle and 3-AFC methods between either sample type. The proportion of correct responses for dilution and compensation samples were larger (71.7% and 78.1%, respectively) when the 3-AFC method was used than the proportion of correct responses for the triangle test method (47.7% and 57.3%, respectively). These results confirm Thurstonian modeling predictions (Frijters 1979).

As discussed by Ennis and Mullen (1985), correlation between sample dimensions may provide additional perceptual information which improves subject performance. Compensation samples resulted in a larger proportion of correct responses and higher power than dilution samples for the triangle test method (Table 6.5). Higher proportion of correct responses for compensation samples suggests that while citric acid was added to compensate for the reduction in lemon-lime flavor the overall formulation change between samples increased the perceived sensory difference. The reduction in sucrose and increase in citric acid used between compensation samples may have worked in conjunction to increase perception of sourness and or decrease the perception of sweetness. Overall, the impact of sweetness and sourness perceptions resulted in a larger d' even when a compensation is made for the reduction of ingredients.

6.5 Conclusions

The results from the current study reveal an impact on proportion of correct responses due to sample dimensionality. While these findings are consistent for both methods tested, the effect of dimensionality and correlation between dimensions may be specific to the sample formulations used in the study. As sensory discrimination testing methods differ in the optimal mind strategy used in selecting an answer, modalities or even specific attributes of a product may induce method-dependent interactions which create differences between the results of specified and non-specified discrimination testing when using multidimensional samples. Differing sample dimensions may impact one method more than another.

While the overall change in formulation between compensation samples led to an increase in the proportion of correct responses compared to dilution samples, in an industrial setting the typical goal would be to reduce the difference between samples. As sucrose is removed from a product typically another sweetener would be utilized to compensate for the reduction in perceived sweetness such as in diet sodas. While sweetness decreases with the reduction of sugar it increases with the addition of artificial sweetener and ideally reduces the perceived difference between full and reduced calorie soft drinks. This situation may lead to a reversal of the findings found in Experiment II, which demonstrates a limitation to the selection of formulations used in studying discrimination methods. While dilution of samples is a convenient mode of creating complex changes between samples, it does not encompass the scope of changes made in the food industry.

Experiment I examined the impact of added dimensions where the only difference between formulations was sucrose. Future research should explore the dependency of the proportion of correct responses on the attribute of difference between samples. It is unclear if

citric acid, flavor, carbonation, or other common beverage ingredients would result in equivalent decreases in proportion of correct responses. Odor and taste sensations were explored within the sample sets tested here, but sample properties such as texture, temperature, and other feeling factors may also cause sample dimensionality and influence the results of sensory discrimination testing. While multidimensional models exist which predict the impacts of dimensionality on discrimination testing, these models require knowledge of the sensory characteristics and their correlations to one another in order to use effectively. In discrimination test situations, these properties may not always be known. Conducting discrimination testing in conjunction with descriptive analysis profiles of the samples may allow for the use of the multidimensional models and lead to increased understanding of how complexity impacts discrimination test results.

6.6 References

Bi J, Ennis DM. 1998. A Thurstonian Variant Of The Beta-Binomial Model For Replicated Difference Tests. J.Sens.Stud. 13(4):461-6.

Bi J, Ennis DM, O'Mahony M. 1997. How to estimate and use the variance of d' from difference tests. J.Sens.Stud. 12(2):87-104.

Bloom DJ. 2015. Sensory Discrimination Testing Methodology Selection Based on Beverage Complexity.

Byer AJ, Abrams D. 1953. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Dalton P, Doolittle N, Nagata H, Breslin PAS. 2000. The merging of the senses: Integration of subthreshold taste and smell. Nat.Neurosci. 3(5):431-2.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM, Mullen K. 1986a. A multivariate model for discrimination methods. J.Math.Psychol. 30(2):206-19.

Ennis DM, Rousseau B, Ennis JM. 2013. Tools and Applications of Sensory and Consumer Science. 3.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Ennis DM, Bi J. 1998. The Beta-binomial Model: Accounting For Inter-trial Variation In Replicated Difference And Preference Tests. J.Sens.Stud. 13(4):389-412.

Ennis DM, Mullen K. 1986b. A multivariate model for discrimination methods. J.Math.Psychol. 30(2):206-19.

Ennis DM, Mullen K. 1986c. Theoretical aspects of sensory discrimination. Chemical Senses 11(4):513-22.

Ennis DM, Mullen K. 1985. The effect of dimensionality on results from the triangular method. Chemical Senses 10(4):605-8.

Frank RA, van der Klaauw NJ, Schifferstein HNJ. 1993. Both perceptual and conceptual factors influence taste-odor and taste-taste interactions. Percept.Psychophys. 54(3):343-54.

Frijters JER. 1979. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER, Kooistra A, Vereijken PFG. 1980. Tables of d' for the triangular method and the 3-AFC signal detection procedure. Percept.Psychophys. 27(2):176-8.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Qual.Preference 31(1):49-55.

Lawless HT, Heymann H. 1999. Sensory Evaluation of Food. Gaithersburg, Maryland: Aspen Publishers, Inc. 827 p.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Masuoka S, Hatjopoulos D, O'Mahony M. 1995. Beer Bitterness Detection: Testing Thurstonian And Sequential Sensitivity Analysis Models For Triad And Tetrad Methods. J.Sens.Stud. 10(3):295-306.

Mata-Garcia M, Angulo O, O'Mahony M. 2007. On Warm-up. J.Sens.Stud. 22(2):187-93.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

O'Mahony M, Odbert N. 1985. A Comparison of Sensory Difference Testing Procedures: Sequential Sensitivity Analysis and Aspects of Taste Adaptation. J.Food Sci. 50(4):1055-8.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M, Rousseau B. 2003. Discrimination testing: a few ideas, old and new. Food Quality and Preference 14(2):157-64.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivit of the same-different test: Comparison with triangle and duo-trio methods. J.Sens.Stud. 13(2):149-73.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-84.

The soft drinks companion: a technical handbook for the beverage industry [Internet]. Boca Raton (Florida) [etc.]: CRC Press; 2005 [Accessed]

Small DM, Prescott J. 2005. Odor/taste integration and the perception of flavor. Experimental Brain Research 166(3):345-57.

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

6.7 Tables and Figures

Table 6.1. Composition of samples created for Experiment 1. Each control and variant pair was included in 3-AFC and triangle testing conditions. The change only occurred in one ingredient, sucrose.

One Dimensional Sample		Two Dimensional Sample		Three Dimensional Sample	
Control	Variant	Control	Variant	Control	Variant
100.0 g Sucrose/L	120.0 g Sucrose/L	100.0 g Sucrose/L 1.00 g Citric Acid/L	120.0 g Sucrose/L 1.00 g Citric Acid/L	100.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L	120.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L

Table 6.2. Composition of samples created for Experiment II. Each control and variant pair was included in 3-AFC and triangle testing conditions.

Dilution		Compensation	
Control	Variant	Control	Variant
100.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L	87.5 g Sucrose/L 0.88 g Citric Acid/L 0.22 g Flavor/L	100.0 g Sucrose/L 0.88 g Citric Acid/L 0.25 g Flavor/L	87.5 g Sucrose/L 1.00 g Citric Acid/L 0.22 g Flavor/L

Table 6.3. Summary of results from Experiment I comparing the results of triangle and 3-AFC procedures with increasing levels of dimensionality. P value listed indicates the significance at which the test method can declare the sample pair different. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by beta binomial calculations. σ^2 indicates the variance of d'. Power is the probability of declaring a difference when one exists between samples. γ is the overdispersion present due to replicated difference testing. 3-AFC procedures were performed with an n=23. Triangle procedures were performed with an n=24. Each subject performed four replications of each test method and dimensionality. One dimensional samples consisted of sucrose in water. Two dimensional samples consisted of sucrose and citric acid in water. Three dimensional samples consisted of sucrose, citric acid, and flavor in water.

	3-AFC			Triangle		
	1 Dimensional	2 Dimensional	3 Dimensional	1 Dimensional	2 Dimensional	3 Dimensional
p	0.00	0.00	0.00	0.00	0.02	0.01
d'	1.90	1.51	1.81	2.16	1.29	1.50
σ^2	0.05	0.08	0.08	0.07	0.12	0.12
Power	1.00	1.00	1.00	1.00	0.70	0.83
γ	0.01	0.45	0.30	0.01	0.18	0.25

Table 6.4. Comparison of experimental values of proportion of correct responses from triangle testing conducted in Experiment I to theoretical predictions made by Ennis and Mullen 1986. Experimental responses were collected from a panel of 24 subjects which performed 4 replicated triangle tests for each sample dimension.

	Theoretical Pc	Experimental Pc
Number of dimensions	d' = 2.2	
1	0.641	0.635
2	0.604	0.467
3	0.575	0.507

Figure 6.1. A comparison of the impact of dimensionality on the proportion of correct responses between 3-AFC (n=23) and triangle (n=24) test methods. One dimensional samples consisted of sucrose in water. Two dimensional samples consisted of sucrose and citric acid in water. Three dimensional samples consisted of sucrose, citric acid, and flavor in water.

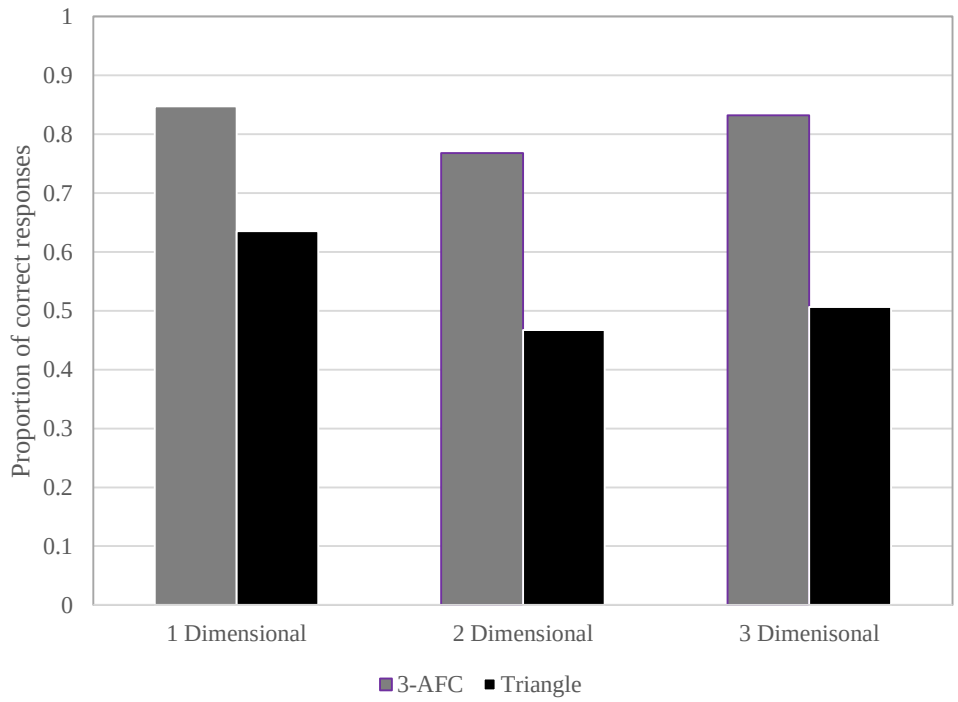


Figure 6.2. A comparison of the impact of dimensionality on d' for 3-AFC (n=23) and triangle (n=24) test methods. Subjects performed four replicated discrimination tests. d' was determined using tables derived by Ennis (1993). Sample comparisons with same letters are not significantly different (α=0.05) within each test method. No significant differences were observed between test methods using the same sample comparisons. One dimensional samples consisted of sucrose in water. Two dimensional samples consisted of sucrose and citric acid in water. Three dimensional samples consisted of sucrose, citric acid, and flavor in water.

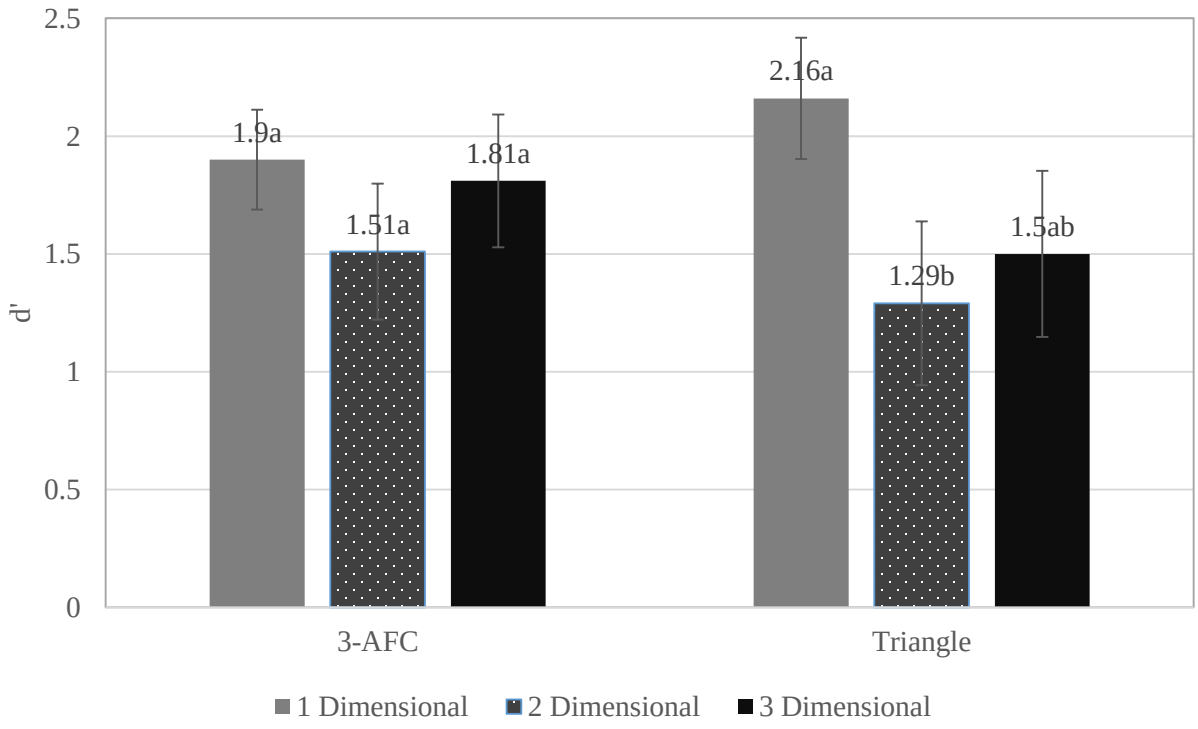


Table 6.5. Summary of results from Experiment II comparing the results of triangle and 3-AFC procedures with increasing levels of dimensionality. P value listed indicates the significance at which the test method can declare the sample pair different. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by beta binomial calculations. σ^2 indicates the variance of d'. Power is the probability of declaring a difference when one exists between samples. γ is the overdispersion present due to replicated difference testing. 3-AFC procedures were performed with an n=23. Triangle procedures were performed with an n=24. Each subject performed four replications of each test method and dimensionality.

	3-AFC		Triangle	
	Dilution	Compensation	Dilution	Compensation
p	0.00	0.00	0.01	0.00
d'	1.30	1.57	1.34	1.84
σ^2	0.05	0.09	0.10	0.07
Power	1.00	1.00	0.80	1.00
γ	0.14	0.50	0.11	0.01

Chapter 7: Impact of d' and Dimensionality on Sensory Discrimination Method Power

7.1 Abstract

Performance on sensory discrimination testing can be impacted by the degree of difference between samples and the dimensionality of the samples used in testing. How these two factors interact and alter performance of discrimination methods is still unknown. The goal of the study was to determine the impact of d' and sample dimensionality interactions on the power of 3-AFC and triangle test method.

Replicated triangle and 3-AFC tests were performed by 47 pre-screened subjects. A three-dimensional model beverage was developed as a base formulation for testing. Samples consisted of sucrose, citric acid, and flavor mixed in filtered water. The base formulation was altered in both unidimensional and multidimensional sample differences to create higher d' and lower d' sample sets. At the completion of each testing session, subjects were asked a multiple choice question to which they answered the strategy used to discriminated the samples. Power, d' and the proportion of correct responses were calculated for each test method.

Overall, the 3-AFC procedures maintained a larger proportion of correct responses and power throughout testing. The triangle test method resulted in larger separation in d' for lower and higher d' samples with both unidimensional and multidimensional formulation than the 3-AFC method. When utilizing multidimensional samples, the 3-AFC method had no significant difference in d' between samples formulated with smaller ingredient differences and larger ingredient differences. Based on these findings, sample dimensionality and d' impact the results of discrimination testing differently between test methods. Future research should explore the impacts of dimensionality over a wider range of d'.

Keywords: Discrimination testing, d', sample dimensionality, Triangle, 3-AFC, Power

7.2 Introduction

Professionals within the food industry use discrimination testing to optimize cost, investigate customer complaints, determine shelf life, and qualify standards for use in other sensory methods. Literature has proven that the methods used to perform sensory discrimination testing can greatly impact subject performance and test power (Byer and Abrams 1953b; Frijters 1979a; Ennis 1990; Ennis 1993; Bi and Ennis 1999; McClure and Lawless 2010). By increasing subject performance in discrimination testing, resources such as the number of subjects needed on a test, the amount of samples which need to be prepared, and the amount of time and other resources required to run the test can be greatly reduced; thus, saving time and money (O'Mahony and Rousseau 2003).

The power of a test method is often dependent on the mind strategy subjects utilize when completing the test. For example, a subject who is asked to perform a triangle test will be presented with three samples and compare them to one another. Samples which are perceived to be closer together on a perceptual continuum, or more similar, will be designated the same and the odd sample identified. This strategy is called the "comparison of distances strategy". A subject completing a 3-AFC test is believed to utilize a "skimming strategy" when completing the testing. Using the skimming strategy a subject will identify the test sample which has the highest perception of the specified attribute regardless of the approximate distance between samples (O'Mahony and others 1994).

Studies conducted to compare discrimination testing methods generally utilize samples which fall into three main categories 1) model solutions (Byer and Abrams 1953a; O'Mahony

and Odbert 1985; Tedja and others 1994; Mata-Garcia and others 2007), 2) samples which are altered with one ingredient (Stillman 1993; Masuoka and others 1995; Delwiche and O'Mahony 1996; Rousseau and O'Mahony 1997; Rousseau and others 1998), or 3) samples with complex changes created using dilution (McClure and Lawless 2010; Ishii and others 2014). Little research has been conducted to study samples which are multidimensional and differ in more than one ingredient change. While dilution is one mode of testing the impact of complex sample changes, ingredient changes in the food industry are rarely dilution based and rarely involve a single ingredient.

One reason why dimensionality is rarely discussed within the literature is basic Thurstonian model's dependence on the assumption that perceptual differences lie along a unidimensional continuum (Frijters 1979b). Multidimensional Thurstonian models have been developed to accommodate multidimensional changes (Ennis and Mullen 1985; Ennis and Mullen 1986a; Ennis and Mullen 1986b; Mullen and Ennis 1987), but without the knowledge of how perceptions are correlated the models become difficult to use or have been deemed unnecessary (Ennis 1998).

In addition to the lack of research available on sample dimensionality, studies which focus on discrimination testing have typically been performed using samples with a large degree of difference between sample pairs (Ishii and others 2014). A way of determining different samples are in terms of perceptual distance is through the estimation of d'. Samples with a higher d' have a greater distance between perceptual means than samples with low d' (O'Mahony and others 1994), and would thus be more distinguishable from each other. The d' of samples typically utilized in literature tend to be beyond those which would typically be perceived as confusable. Discrimination testing aims to identify if there is a perceivable difference between

samples. Samples which are not confusable will obviously be found to be different when analyzed using a discrimination test. Thus, these samples which are not confusably different would not deem suitable for discrimination testing for a sensory scientist within the food industry.

Two experiments were conducted to explore the influence of sample d' and multidimensionality on the results of specified 3-AFC and non-specified triangle test methods. In Experiment I, the impact of sample set d' on the proportion of correct responses and power for Triangle and 3-AFC testing methods was assessed. It was hypothesized that specified test methods will have a higher proportion of correct responses than non-specified methods when multidimensional samples with higher d' are used based on literature findings and theoretical predictions discussed previously. The goal of Experiment II was to determine if d' impacts method power in samples with multidimensional formulations which differ in multiple dimensions simultaneously. It was hypothesized that at lower d', sample dimensionality will cause increased interstimulus variation and decrease the proportion of correct responses for triangle and 3-AFC methods.

7.3 Materials and Methods

7.3.1 Experiment I

7.3.1.1 Subjects

All subjects who participated in testing were recruited from the University of Illinois at Urbana-Champaign campus. Subjects were recruited and screened to ensure a base level of sensory acuity across subjects. In order to participate in the study subjects needed to obtain a minimum of 70% correct responses and be free from food allergies. Screening procedures for the

study included taste identification and intensity ranking, aroma identification, performance on discrimination testing methods including the triangle and duo-trio methods, as well demographic information. Detailed information on the samples used in prescreening procedures can be found in Chapter 4 (Bloom 2015). 69% of subjects screened for the study obtained the minimum of 70% correct responses and were included in the study.

All subjects were between the ages of 18 and 55 years of age; 32 were female and 15 were male. Subjects were randomly assigned to either 3-AFC or triangle test groups. Twenty-four subjects began testing in each test group, but shortly after beginning, one of the subjects from the 3-AFC test group dropped from the study which accounts for the differences in sample size for the two test groups. At the completion of testing, subjects were compensated monetarily. Subjects utilized in Experiment I also completed testing in Experiment II.

7.3.1.2 Samples

Samples used in testing were created using an example drink formulation found in Maurice Shachman's The Soft Drinks Companion (Shachman 2005) and were modified to fit the needs of the current study. Higher and lower d' sample sets were created to determine the impact of d' on subject performance in multidimensional samples. Samples consisted of reverse osmosis filtered water with three added dimensions: sucrose, citric acid, and lemon-lime flavor. Samples used in testing differed in the amount of sucrose used in formulation. For higher d' samples, the difference in sucrose used between test pairs was larger than that found in lower d' samples. Sample formulations can be found in Table 7.1. Samples were served in lidded, 60 mL clear plastic cups (Dart Container Corporation, Mason, MI) which were labeled with random three-digit codes.

Preliminary testing was conducted in order to determine the d' of samples used in testing. Samples were created so that higher d' samples were slightly below a d' of 2.0. For each test pair comparison, 15 subjects performed four replicate triangle tests. A total of 20 preliminary testing sessions were conducted. Several subjects participated in multiple preliminary tests. A total of 32 subjects (21 female, 11 male) performed preliminary testing.

7.3.1.3 Experimental Procedure

3-AFC and triangle test methods were utilized in the completion of Experiment I. Subjects were randomly assigned to one of the two test methods and performed the same method throughout the experiment. A between groups design was employed in order to prevent subjects from changing decision strategies when completing 3-AFC or triangle methods and to limit fatigue (Rousseau and others 1999; McClure and Lawless 2010). A detailed screening procedures was completed by each subject to ensure a base level of sensory acuity and decrease variation in sensitivity of subjects within each test group. Each subject completed two 30-minute test sessions. In each session subjects performed four replicate discrimination tests with a 2-minute break between each test. Higher and lower d' samples were tested in different sessions for each subject. The order of test sessions and sample presentation order was randomized across subjects. For 3-AFC procedures, subjects were instructed to identify the sweeter sample as sample differed in the amount of sucrose present.

Subjects were instructed to rinse their mouths before beginning testing and between each sample with warm water (43-49°C) followed by room temperature water (22°C). Testing was conducted in isolated sensory booths. Booths were temperature controlled at 22°C and a relative

humidity of 33%. Data was collected using Compusense® *five Plus* (Version 5.6: Guelph ON, Canada).

7.3.1.4 Data Analysis

Data were analyzed using IFPrograms™ version 8.12 (The Institute for Perception, USA). As replications were performed, beta binomial analysis was conducted to account for variability between samples and subjects (Liggett and Delwiche 2005). The d' of sample pairs was determine using tables derived by Ennis (1993). The variance of d' was calculated using methods described by Bi and others (1997). Power for each test and product combination was calculated using the calculated d', an alpha of 0.05, the respective number of trials and replications and calculated gamma values using the IFPrograms™ software. Chi-Square analysis was performed to determine if significant differences existed between d' values produced from different samples and test methods.

7.3.2 Experiment II

7.3.2.1 Subjects

Subjects who participated in Experiment I also participated in Experiment II. Subjects were assigned to the same test group for both experiments.

7.3.2.2 Samples

Multidimensional samples were created using the same sample ingredients found in Experiment I. Sample formulations for Experiment II can be found in Table 7.2. Multiple ingredients were varied between control and variant formulations used in the study. As sucrose

was the largest ingredient change between control and variant, in 3-AFC procedures subjects were asked to identify the sweeter sample.

The formulation of samples was determined through preliminary testing. Samples were created in order to achieve a d' below 1.0 for low d' and approximately 2.0 for high d' samples. Pretesting was conducted through the recruitment of 15 subjects for each product comparison. Each subject completed four replicate triangle tests. Data were analyzed to determine an estimation of d'.

7.3.2.3 Experimental Procedure

A total of two test sessions was completed by each subject. Only one sample type, either high or low d', was tested in each testing session. The order of sessions was randomized across subjects. Subjects performed four replicate discrimination tests in each testing session with a two-minute break between each replicate. Subjects were instructed to rinse with warm water (43-49°C) and room temperature water (22°C) prior to beginning testing and between each test sample. Testing was completed in sensory booths where the temperature was maintained at 22°C. Data were collected using Compusense® *five Plus* (Version 5.6: Guelph ON, Canada).

At the completion of each session, subjects were asked to identify how they discriminated samples in testing using a multiple choice question. Test answers were selected to represent the two possible mind strategies employed when completing specified and non-specified test methods. Presentation of multiple choice answers was randomized throughout testing.

7.3.2.4 Data Analysis

Beta binomial analysis and power calculations were conducted using IFPrograms™ version 8.12 (The Institute for Perception, USA). Power was determined using the calculated overdispersion, four replications, number of subjects (23 for 3-AFC, 24 for Triangle), alpha of 0.05, and a null probability of 1/3. Chi-square analysis was performed to determine if d' values from different test methods and d' levels were significantly different.

7.4 Results and Discussion

7.4.1 Experiment I

A summary of results for Experiment I can be found in Table 7.3. For both higher and lower d' samples the 3-AFC test method resulted in a higher proportion of correct responses (high d' = 82.4%, low d' = 72.4%) than the triangle test method (high d' = 59.5%, low d' = 43.8%). These results confirm Thurstonian modeling predictions and confirm the paradox of discriminatory nondiscriminators (Frijters 1979a). The d' values associated with each product comparison do not differ significantly (alpha=0.05) between test methods.

Samples used in Experiment I differed in the amount of sugar between control and variant pairs and represented a unidimensional change between samples. Lower d' samples had a smaller difference in sucrose than higher d' samples. It was expected that samples with smaller formulation differences would result lower d' values than samples with larger formulation differences. When comparing the d' values associated with higher and lower d' samples (Table 7.3), no significant differences exist for the 3-AFC method. The results from the triangle test method did lead to significant differences (p=0.048) between the d' values of higher and lower d' samples.

While the proportion of correct responses is lower for the triangle test method than the 3-AFC test method, the triangle method is measuring a larger difference between low d' and high d' samples. Samples used in testing differed based on the amount of sucrose used in formulation, but the perceptual change of sucrose may also have been impacted by the presence of other compounds such as flavor (Frank and others 1993). The 3-AFC method instructed subjects to identify the sweeter sample. It is possible that the triangle test method allowed subjects to integrate differences caused by taste and flavor perceptions and thus resulted in larger differences between high d' and low d' samples.

Both triangle and 3-AFC methods found significant differences between control and variant samples (Table 7.3). The 3-AFC method was found to be more powerful than the triangle test method when lower d' samples were compared. At higher levels of d' both triangle and 3-AFC test methods reached a power of 1.0 as observed in table 7.3. These results demonstrate a ceiling effect, which is reached when using samples with high d'. While the proportion of correct responses is larger for the 3-AFC method, both methods are equally powerful. These results suggest that in order to observe true power differences between test methods, samples with lower d' should be used in testing. At lower d', the triangle test method resulted in the reduction of power from 1.0 to 0.68 with the reduction in d' between samples (Table 7.3). The 3-AFC method maintained a high level of power for both high and low d' samples.

As subjects performed replicated tests, overdispersion (Cox 1983; Anderson 1988) has been estimated to represent the variation caused by subjects (Ennis and Bi 1998) and is expressed as γ in Table 7.3. For both lower and higher d' samples, overdispersion is greater for the 3-AFC method than for the triangle method. While not found to be significant, Liggett and Delwiche (2005) found a similar trend when comparing methods using cherry-flavored

beverages. When considering the effects that sample sequence has on discriminability of samples, it would be expected that the 3-AFC method, having two weaker samples and one stronger sample, would have lower variation than the triangle test which includes presentations where two stronger and one weaker sample (Vie and O'Mahony 1989; Liggett and Delwiche 2005). Therefore, sequential sensitivity does not explain the increased variation observed for the 3-AFC method. One possible explanation for the greater overdispersion of the 3-AFC method is the fact that subjects were different between test groups. Further explored in Experiment II, subjects may utilize decision strategies other than the theorized strategy, which may increase subject variation and create inconsistencies with relating results to model predictions.

7.4.2 Experiment II

A summary of results from Experiment II can be found in Table 7.4. Thurstonian modeling confirmed as the 3-AFC test method resulted in a larger proportion of correct responses for both d' levels (high d' = 80%, low d' = 81.5%) than the triangle test method (high d' = 56.2%, low d' = 47.9%).

Samples used in testing differed in several dimensions. Sucrose and flavor levels decreased between control and variant pairs while citric acid levels increased. The expected result of these formulation changes was a difference in sweetness and sourness between samples used in testing. As in experiment I, samples with smaller formulation differences were expected to result in lower d' values than samples with larger formulation differences. When comparing test methods and samples used in testing, no significant differences were found between d' values in the study (alpha=0.05). Contrary to expectations, subjects using the 3-AFC method had a higher proportion of correct responses when samples with smaller formulation differences were

used in testing compared to samples with larger formulation differences. As in Experiment I, there is a greater difference between the d' values for high and low d' samples with the triangle test method than there is for the 3-AFC test method. The triangle test method follows expected results based on formulation differences more closely than the 3-AFC test method when multidimensional changes are made between samples. While this may be helpful in a product development setting, the lower power would necessitate additional subjects to obtain high test power.

At the completion of each test session, subjects were asked to identify how they differentiated samples on a multiple choice question. The percentage of responses for each choice selection can be found in Table 7.5. Based on theory predictions, subjects are assumed to employ a skimming decision strategy when completing the 3-AFC method and a comparison-of-distances strategy when performing the triangle method (O'Mahony and others 1994). Based on responses to the multiple choice question used in Experiment II, 43.5% of subjects utilized a comparison-of-distances strategy when completing the specified 3-AFC test method. When combining the frequency of responses for choices indicating a skimming strategy, 20-29% of subjects completed the triangle test method utilized the skimming strategy. These results demonstrate that there is not a complete uniformity of decision strategy when completing discrimination testing methods. Additionally, the comparison of distances strategy was found to commonly be utilized even when an attribute is provided to subjects for use in specified discrimination testing.

As in Experiment I, overdispersion was higher for the 3-AFC method than the triangle method (Table 7.4). For the 3-AFC method overdispersion increased as formulation differences between samples decreased. The higher proportion of subjects using a theorized decision strategy

may be a possible explanation for the higher overdispersion for the 3-AFC method. There is greater variability in the decision strategy which subjects identified as using for the 3-AFC method than the triangle method, so one may assume that the increased variation in decision strategy has led to higher overall subject variation.

7.5 Conclusions

The present research expands upon the knowledge available in the area of discrimination testing methodologies and present findings relevant to the types of samples and number of subjects commonly used in the context of commercial food products. The research provides insight as to how sample complexity and degree of difference impact the expected results. One possible limitation to the findings is the range of d' used in testing. Lower d' samples may be above the range of samples common to those used in the food industry, but are relevant to the range of samples used within literature.

While the findings provided here have shown differences in triangle and 3-AFC method results using multidimensional samples at two levels of d', these findings should be explored over a larger range of d'. Further research should be conducted using samples with a d' below 1.0 to determine the impact of both sample dimensionality and level of d' on sensory discrimination testing power.

As the current study found relatively large proportions of subjects utilizing a decision strategy other than the theorized strategy, monitoring of sensory panel decision strategies should be adopted to determine if adjustments need to be made to models. Findings suggest that method training may need to be employed to increase the proportion of subjects who utilize the theorized decision strategy. Consider an analogy to tennis. As a player practices proper technique of

serving and studies the physical mechanics which have been demonstrated as successful, a player may improve in their performance during a match. The player may revert to old habits occasionally, but overall performance is improved. Applying this to discrimination testing, while the theorized decision strategy may not be utilized consistently by all subjects, identifying the decision strategy which leads to greater chance of success may decrease the proportion of incorrect responses. Discrimination testing conducted using naïve consumers may have a larger proportion of subjects utilizing non-theorized decisions strategies, and may, thus, differ from a trained panel. Variability in sensory discrimination testing can be found in both samples and subjects. Exploring the effects of sample dimensionality and degree of difference on discrimination method power has and will continue to provide guidance to sensory professionals in the selection of appropriate methods for the study they conduct.

7.6 References

Anderson DA. 1988. Some models for overdispersed binomial data. Australian Journal of Statistics 30(2):125-48.

Bi J, Ennis DM. 1999. The Power Of Sensory Discrimination Methods Used In Replicated Difference And Preference Tests. J.Sens.Stud. 14(3):289-302.

Bi J, Ennis DM, O'Mahony M. 1997. How to estimate and use the variance of d' from difference tests. J.Sens.Stud. 12(2):87-104.

Byer AJ, Abrams D. 1953a. A comparison of the triangular and two-sample taste-test methods. Food Technol. 7185.

Byer AJ, Abrams D. 1953b. A comparison of the triangular and two-sample taste-test methods. Wallerstein Lab Commun 16((54)):253-60.

Cox DR. 1983. Some remarks on overdispersion. Biometrika 70(1):269-74.

Delwiche J, O'Mahony M. 1996. Flavour discrimination: An extension of thurstonian 'Paradoxes' to the tetrad method. Food Quality and Preference 7(1):1-5.

Ennis D. 1990. Relative Power of Difference Testing Methods in Sensory Evaluation. Food Technol. 44(4):114.

Ennis DM. 1998. Thurstonian Scaling for Difference Tests. IFPress 1(3):2.

Ennis DM, Mullen K. 1986a. A multivariate model for discrimination methods. J.Math.Psychol. 30(2):206-19.

Ennis DM. 1993. The Power Of Sensory Discrimination Methods. J.Sens.Stud. 8(4):353-70.

Ennis DM, Bi J. 1998. The Beta-binomial Model: Accounting For Inter-trial Variation In Replicated Difference And Preference Tests. J.Sens.Stud. 13(4):389-412.

Ennis DM, Mullen K. 1986b. Theoretical aspects of sensory discrimination. Chemical Senses 11(4):513-22.

Ennis DM, Mullen K. 1985. The effect of dimensionality on results from the triangular method. Chemical Senses 10(4):605-8.

Frank RA, van der Klaauw NJ, Schifferstein HNJ. 1993. Both perceptual and conceptual factors influence taste-odor and taste-taste interactions. Percept.Psychophys. 54(3):343-54.

Frijters JER. 1979a. The paradox of discriminatory nondiscriminators resolved. Chemical Senses and Flavour 4(4):355.

Frijters JER. 1979b. Variations of the triangular method and the relationship of its unidimensional probabilistic models to three-alternative forced-choice signal detection theory models. Br.J.Math.Stat.Psychol. 32(2):229-41.

Ishii R, O'Mahony M, Rousseau B. 2014. Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. Food Qual.Preference 31(1):49-55.

Liggett RE, Delwiche JF. 2005. The Beta-binomial Model: Variability In Overdispersion Across Methods And Over Time. J.Sens.Stud. 20(1):48-61.

Masuoka S, Hatjopoulos D, O'Mahony M. 1995. Beer Bitterness Detection: Testing Thurstonian And Sequential Sensitivity Analysis Models For Triad And Tetrad Methods. J.Sens.Stud. 10(3):295-306.

Mata-Garcia M, Angulo O, O'Mahony M. 2007. On Warm-up. J.Sens.Stud. 22(2):187-93.

McClure S, Lawless HT. 2010. Comparison of the triangle and a self-defined two alternative forced choice test. Food Quality & Preference 21(5):547-52.

Mullen K, Ennis DM. 1987. Mathematical formulation of multivariate euclidean models for discrimination methods. Psychometrika 52(2):235-49.

O'Mahony M, Odbert N. 1985. A Comparison of Sensory Difference Testing Procedures: Sequential Sensitivity Analysis and Aspects of Taste Adaptation. J.Food Sci. 50(4):1055-8.

O'Mahony M, Masuoka S, Ishii R. 1994. A Theoretical Note On Difference Tests: Models, Paradoxes And Cognitive Strategies. J.Sens.Stud. 9(3):247-72.

O'Mahony M, Rousseau B. 2003. Discrimination testing: a few ideas, old and new. Food Quality and Preference 14(2):157-64.

Rousseau B, O'Mahony M. 1997. Sensory Difference Tests: Thurstonian And SSA Predictions For Vanilla Flavored Yogurts. J.Sens.Stud. 12(2):127-46.

Rousseau B, Meyer A, O'Mahony M. 1998. Power and sensitivit of the same-different test: Comparison with triangle and duo-trio methods. J.Sens.Stud. 13(2):149-73.

Rousseau B, Rogeaux M, O'Mahony M. 1999. Mustard discrimination by same–different and triangle tests: aspects of irritation, memory and τ criteria. Food Quality and Preference 10(3):173-84.

The soft drinks companion: a technical handbook for the beverage industry [Internet]. Boca Raton (Florida) [etc.]: CRC Press; 2005 [Accessed]

Stillman JA. 1993. Response selection, sensitivity, and taste-test performance. Percept.Psychophys. 54(2):190-4.

Tedja S, Nonaka R, Ennis DM, O'Mahony M. 1994. Triadic discrimination testing: Refinement of Thurstonian and sequential sensitivity analysis approaches. Chem.Senses 19(4):279-301.

Vie A, O'Mahony M. 1989. Triangular difference testing: refinements to sequential sensitivity analysis for predictions for individual triads. J.Sens.Stud. 4(2):87-103.

7.7 Tables and Figures

Table 7.1. Composition of samples created for Experiment 1. Each control and variant pair was included in 3-AFC and triangle testing conditions.

Higher d’		Lower d’	
Control	Variant	Control	Variant
100.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L	127.5 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L	100.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L	110.0 g Sucrose/L 1.00 g Citric Acid/L 0.25 g Flavor/L

Table 7.2. Composition of samples created for Experiment II. Each control and variant pair was included in 3-AFC and triangle testing conditions.

High d’		Low d’	
Control	Variant	Control	Variant
100.0 g Sucrose/L 0.85 g Citric Acid/L 0.25 g Flavor/L	85.0g Sucrose/L 1.00 g Citric Acid/L 0.21 g Flavor/L	100.0 g Sucrose/L 0.93 g Citric Acid/L 0.25 g Flavor/L	92.5 g Sucrose/L 1.00 g Citric Acid/L 0.23 g Flavor/L

Table 7.3. Summary of results from Experiment I comparing the results of triangle and 3-AFC procedures with different levels of d'. P-value listed indicates the significance at which the test method can declare the sample pair different. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by beta-binomial calculations. σ^2 indicates the variance of d'. Power is the probability of declaring a difference when one exists between samples. γ is the overdispersion present due to replicated difference testing. 3-AFC procedures were performed with an n=23. Triangle procedures were performed with an n=24. Each subject performed four replications of each test method and dimensionality.

	3-AFC		Triangle	
	Higher d'	Lower d'	Higher d'	Lower d'
p-value	0.00	0.00	0.00	0.02
d'	1.76	1.33	1.95	1.12
σ^2	0.10	0.07	0.08	0.09
Power	1.00	1.00	1.00	0.68
γ	0.47	0.34	0.10	0.01

Table 7.4. Summary of results from Experiment II comparing the results of triangle and 3-AFC procedures with multidimensional changes at high and low d'. P value listed indicates the significance at which the test method can declare the sample pair different. d' is the distance in perceptual standard deviations between sample pairs utilized for each sample and method as determined by beta binomial calculations. σ^2 indicates the variance of d'. Power is the probability of declaring a difference when one exists between samples. γ is the overdispersion present due to replicated difference testing. 3-AFC procedures were performed with an n=23. Triangle procedures were performed with an n=24. Each subject performed four replications of each test method and d' level.

	3-AFC		Triangle	
	High d'	Low d'	High d'	Low d'
p	0.00	0.00	0.00	0.00
d'	1.66	1.72	1.79	1.36
σ^2	0.06	0.08	0.07	0.08
Power	1.00	1.00	1.00	0.90
γ	0.18	0.35	0.01	0.01

Table 7.5. Percentage of subjects indicating each choice selection when asked to identify how they differentiated samples in Experiment II. The total number of subjects in the 3-AFC test group was 23. The total number of subjects in the Triangle test group was 24. Subjects selected answered the multiple choice question at the end of both test sessions (High d' samples and Low d' samples).

Responses to the question, "How did you differentiate samples?"	3-AFC		Triangle	
	High d'	Low d'	High d'	Low d'
I selected sample highest in one attribute	52.2%	56.5%	20.8%	12.5%
I selected the sample lowest in one attribute	4.3%	0.0%	8.3%	8.3%
I found one sample that was most different from the other two samples	43.5%	43.5%	70.8%	79.2%
Other	0.0%	0.0%	0.0%	0.0%

Chapter 8: Future Directions

8.1 Extension of Current Research

There are several future directions which may be proposed from the findings of the research presented in this dissertation. One unexpected finding from Chapter 7 was the larger overdispersion for the 3-AFC method than the triangle method. A follow-up to the research presented would be to determine if the same differences in overdispersion occur when the same subjects perform both methods. Although all subjects were screened for the same level of sensory acuity, it is possible that the subjects performing 3-AFC procedures were prone to more variability than those performing the triangle test. If the same results are found with subjects performing both methods, the findings would indicate that the decision strategy differences between methods impact subject variability when multidimensional stimuli are tested.

The effects of higher and lower d' samples were compared in Chapter 7. Samples were expected to differ in the degree of difference based on the changes in formulation made between samples pairs. Samples with smaller differences in formulations were expected to have smaller d' values. Samples with larger differences in formulations were expected to have larger d' values. The actual results of the study did not follow this trend. The proportion of correct responses for the 3-AFC method resulted in a slightly higher d' with smaller formulation difference between samples than those with larger formulation differences between samples. In contrast, the triangle method resulted in a lower d' for samples with smaller formulation differences between samples. It is assumed that small changes in formulation would result in small d' between samples. If small changes in formulation result in larger d' values than products with larger changes in formulation, this assumption is violated when 3-AFC test is conducted. An issue is, then, created for the product developer as to what course of action should be taken.

The results from the 3-AFC test method may signify significant differences between products but the relevance of the difference to the consumer is unknown. The discrimination test has served the intended purpose of identifying differences between products, but what is the meaning of the difference?

If perceivable difference is looked at in a real world setting we may further relate d' to the consumer. A reformulated product is not commonly advertised as having changed unless the difference between new and old formulations are known and believed to increase purchase. If the difference is unknown to the consumer, it is unclear as to what level of d' could be recognized. It is not common for consumers to have two variations in one setting to compare samples. For this reason, memory needed between sampling is very high and may increase the difference between samples with which subjects would recognize. An extension of this research would be to conduct consumer testing using samples with wide ranging d' to determine the impact of d' on consumer acceptance. Results of consumer testing could be used to guide those in the food industry to the relevance of d' found in discrimination testing.

8.2 Application of Findings to Consumer Testing

8.2.1 Objective

The objective of this research is to determine how context effects created from environment and sample differences impact the results of acceptability testing of beverages.

8.2.2 Background

Sensory discrimination testing enables researchers to identify differences between samples. The research presented in this dissertation allows for the selection of appropriate

discrimination method based on the sample complexity and d'. While differences between products, estimated by d', provide guidance in product decisions, they are not an indicator of consumer liking. Samples which have large differences in perception and formulation may, in fact, have equivalent consumer acceptance. There is no literature which has demonstrated how the degree of difference between products (d') affects the overall liking of the products being tested. A connection between d' and consumer liking would enable product developers to determine the relevance of differences found in discrimination testing to consumer acceptance of products and predict how liking might change due to context effect by the spectrum of samples being tested.

In addition to the work presented in Chapters 3-7 (Bloom 2015), consumer research was conducted to determine how the environment in which consumer testing occurs affects overall liking for beverage products. The testing locations used in the study were a study room, fitness center, food court, restaurant, and sensory lab setting. Commercially available ginger ale, sparkling water, and tea products were tested by consumers in each location. Small differences were observed for the average overall liking scores of the products in different testing environments with the restaurant setting having the most positive impact on overall liking across products (6.53 on 9-point hedonic scale) and the sensory lab setting having the most negative impact on overall liking (6.29 on 9-point hedonic scale).

Chapters 3-7 from this dissertation (Bloom 2015) demonstrated the impact of sample dimensionality and d' on the results of sensory discrimination testing. Sample d' in addition to testing environment may also impact the results of sensory consumer testing. In the consumer testing study, no significant difference in overall liking was observed between the ginger ale products used in testing, but the products were most greatly affected by the changes to testing

environment. It is hypothesized that consumer acceptance of products is impacted by both testing location and degree of difference across samples, estimated by d'. It is unclear if samples with large differences presented in a consumer test setting influence overall liking scores for individual products. The context effect of sample differences in a consumer test setting will be explored in two testing environments, the restaurant setting and the laboratory setting, as they showed the greatest influence on product liking from our previous study.

8.2.3 Experimental Approach

Subjects

One hundred consumers will be recruited to test at two testing locations, the Spice Box (restaurant setting) and the sensory lab (laboratory setting). Subjects will be recruited through email and flyers posted in buildings on the University of Illinois Campus. In order to participate in the study, subject must be at least 18 years of age and frequent consumers of the product categories utilized in the study.

Samples

Samples will consist of commercially available beverages. Three beverage product categories will be used in testing which have wide product variation. These categories are cola, root beer, and ginger ale. Five products will be selected as samples of focus within each product category. Products will be selected to have large and small differences. Small sensory differences will be created by storing products in PET vs aluminum packaging and large differences will be represented by competitor products. For the cola category, Dr Pepper in PET, Dr Pepper stored in aluminum cans, RC, Coca Cola, and Pepsi have been selected to have a range of d'. For the

root beer category, A&W root beet stored in PET, A&W store in aluminum, IBC, Mug, and Barq's will be used. For the ginger ale category, Canada Dry stored in PET, Canada Dry stored in aluminum, Schweppes, Vernors, and Seagrams will be used.

Experimental Procedure

Testing will be performed using the same sample sets in two different testing environments. Samples will be presented to subjects in randomized order. . Three test sessions will take place in each testing environment, one for each product category. Session order, testing environment, and sample order will be randomized across all subjects. Subjects will be asked to rate the samples on a 9-point hedonic scale ranging from 1= dislike extremely to 9= like extremely. Overall liking will be followed by aroma, taste, and mouthfeel attribute liking questions. All responses will be collected using paper ballots. One hundred subjects will participate in three 20-minute testing session at each of the two testing environments for a total of six test sessions.

Testing environments include:

- Sensory laboratory setting: Testing will be performed in individual booth settings with controlled environment. Room conditions will be restricted to incandescent lighting, relative humidity of 33%, and temperature of 70°F
- Restaurant setting: Testing will be performed Bevier Hall Spice Box, a mock fine dining restaurant setting. Panelists will be seated in a group setting at tables and chairs with fluorescent and natural lighting.

Data Analyses

Consumer data will be analyzed using XLSTAT software to determine if significant differences exist between the same products tested in differing testing environments. Additionally, a comparison will be made between the hedonic scores of samples tested in two different sample context settings. Interaction between testing environment and sample context will be analyzed for significant impact to hedonic scores.

8.2.4 Impact of Research

Findings from this study will allow for the understanding of context effects on consumer sensory testing. The insights gained from the results of the study can be utilized to determine a correction factor for consumer acceptance scores based on the difference across the samples as well as location in which testing was conducted.